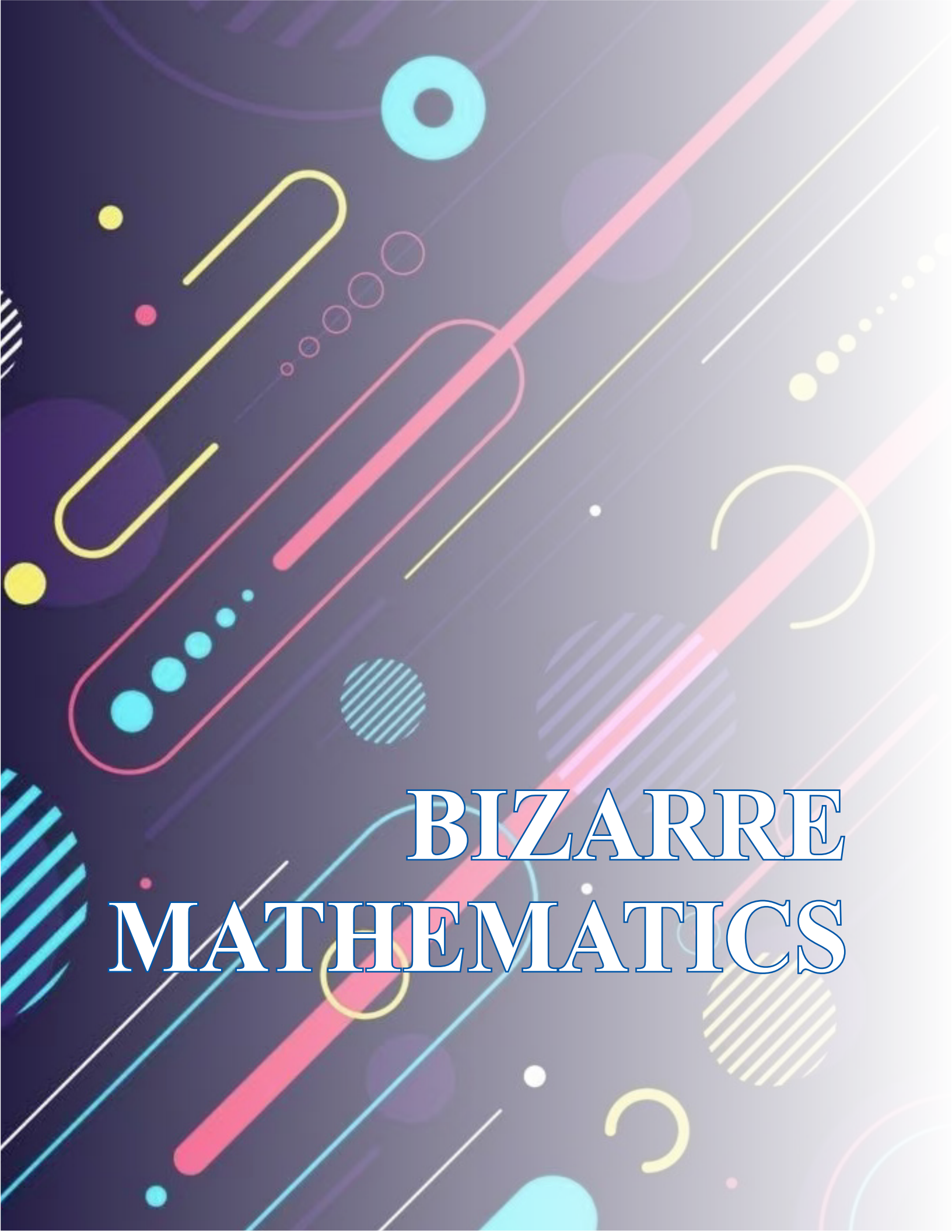

BIZARRE
MATHEMATICS

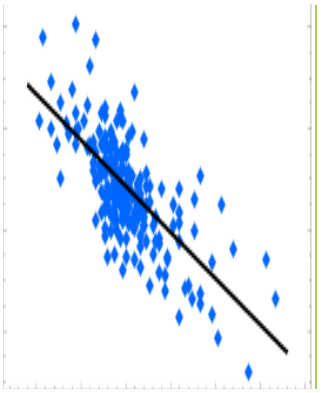
BIZARRE MATHEMATICS

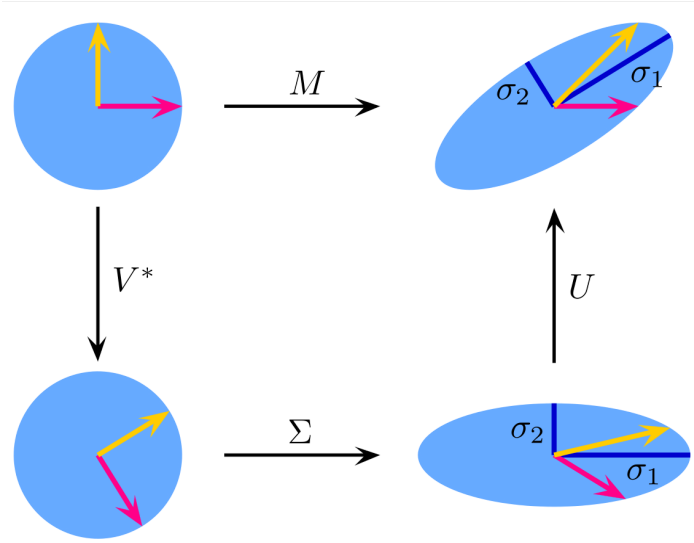
TABLE OF CONTENTS

TABLE OF CONTENTS

MODERN REGRESSION ANALYSIS

SINGULAR VALUE DECOMPOSITION AND REGRESSION





$$M = U \cdot \Sigma \cdot V^*$$

Coveriance Matrix / Singular Value Matrix

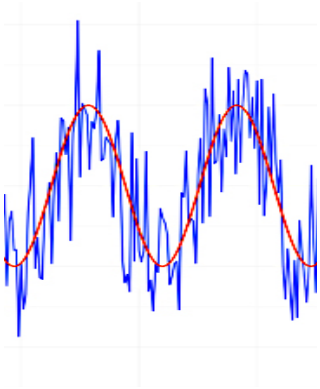
The singular value diagonal matrix is proportional to the diagonal matrix composed of only the diagonal of the variance-covariance matrix in multiple linear regression. The variance-covariance matrix for the design matrix denoted as V, is related to the diagonal singular value matrix, denoted as S, through the following equation:

$$V = \sigma^2(X^T X)^{-1} = \sigma^2 U S^2 U^T ,$$

where σ^2 is the variance of the errors in the regression model, X is the design matrix, U is the matrix of left singular vectors, and S is the diagonal matrix of singular values.

The singular values in S represent the square roots of the eigenvalues of the matrix $X^T X$. They provide information about the amount of variation in the data captured by each of the orthogonal components in the matrix U. The diagonal elements in V, on the other hand, represent the variance of each coefficient estimate in the regression model.

Thus, the relationship between the variance-covariance matrix for the design matrix and the diagonal singular value matrix is that they are related by the variance of the errors in the model and by the decomposition of the design matrix into its orthogonal components, represented by the singular value matrix and the matrix of left singular vectors. ■



MULTIPLE LINEAR REGRESSION AND FOURIER SERIES

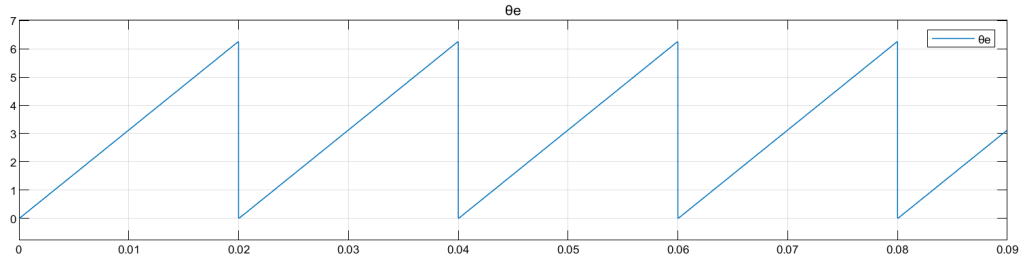
Getting Rid of Noisy Signals

From A Regression POV

Signal distributions are of the form $y(t) = x(t) + \eta(t)$, where *y(t)* is our observed outcome, *x(t)* is the desired distribution, and *η(t)* is the Gaussian white noise.

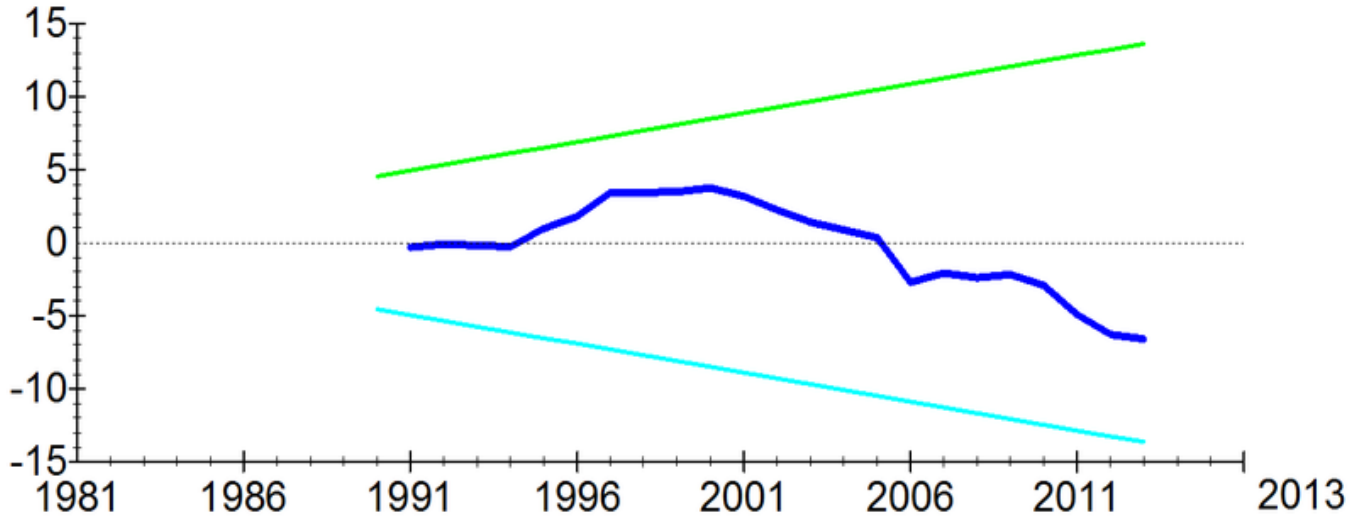
Suppose we want to derive *x(t)* from our observed *y(t)*. In that case, we can first compose our design matrix *X*, which has each column following a sinusoidal distribution with different periods and amplitudes. Then, we can estimate our coefficient matrix β, which has MLE equivalent to $(X^TX)^{-1}X^Ty(t)$.

If *y(t)* is highly correlated with the distribution of a particular column in *X*, then the corresponding coefficient in β for that column will also be high, and vice versa. That way, we can filter out all the unwanted Gaussian noises using linear regression.



Signal filtering and linear regression with a design matrix with each column being a sinusoidal distribution are related. They both involve working with signals that can be decomposed into different frequency components to extract useful information from signals. The focus of using linear regression is identifying and estimating each component's contribution to the response variable, and finally removing the unwanted components. ■

LEAST SQUARES AND THE MINIMIZATION OF RECURSIVE RESIDUALS



Least Squares = Minimization of Recursive Residuals

In multiple linear regression, the maximum likelihood estimate of β can be found using finding the minimum point of the convex least square equation using gradient descent. Usually, we see the process as a process of finding the minimum value of the least squares equation with regard to β. However, the process can also be seen as the minimization of recursive residuals.

Since the fitted least squares line can be seen as a projection of the observed y value to the $\Gamma = X\beta$ hyperplane containing all the values of $\hat{y}$, the errors are naturally the distance between the observed y and $\hat{y}$ in Γ and are orthogonal to every $\hat{y}$. Thus, the errors y_i can be expressed as:

$$e(y_i, \hat{y}_i) = y_i - \text{proj}_\Gamma y_i = y_i - \hat{y}_i \frac{\langle y_i, \hat{y}_i \rangle}{\langle \hat{y}_i, \hat{y}_i \rangle}.$$

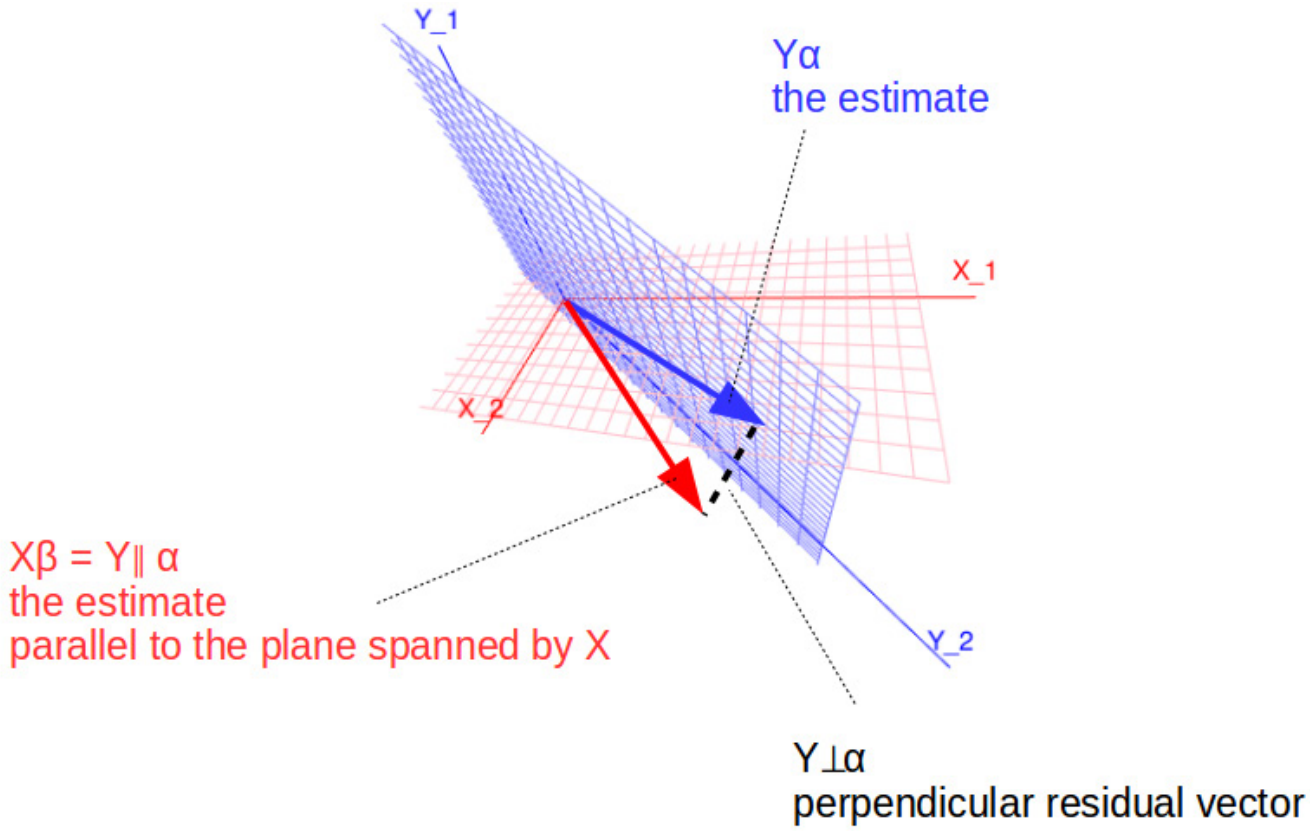
If we write our design matrix X as a matrix of its columns $[X_1, X_2, ..., X_n]$, where each X_i is the i^{th} column of X, then, the least squares equation can be expressed as:

$$||Y - X\beta|| = ||Y - X_1\beta_1 - X_2\beta_2 - ... - X_n\beta_n||^2.$$

We can then choose a particular $X_i\beta_i$ and fix the rest of the equation $||(Y - X_1\beta_1 - ... - X_{i-1}\beta_{i-1} - X_{i+1}\beta_{i+1} - ... - X_n\beta_n) - X_i\beta_i||^2$ to find the MLE of β_i. In other words, we use X_i to predict $(Y - X_1\beta_1 - ... - X_{i-1}\beta_{i-1} - X_{i+1}\beta_{i+1} - ... - X_n\beta_n)$. Without loss of generality, pick

out the column X_1 and its corresponding coefficient β_1. The MLE of β_1 can be found using the formula $\beta_1 = \frac{\langle Y - X_2\beta_2 - \ldots - X_n\beta_n, X_1\rangle}{\langle X_1, X_1\rangle} = \frac{\langle Y, X_1\rangle}{\langle X_1, X_1\rangle} - \beta_2\frac{\langle X_2, X_1\rangle}{\langle X_1, X_1\rangle} - \ldots - \beta_n\frac{\langle X_n, X_1\rangle}{\langle X_1, X_1\rangle}$, since the MLE of any coefficient matrix β is $(X^TX)^{-1}X^TY$. Substituting the value of β_1 into the least squares equation and applying the formula for residuals, we get a partially minimized least square loss function which has the coefficient β_1 minimized to its optimized value. That is,

$$||(Y - X_2\beta_2 - \ldots - X_n\beta_n) - X_1\beta_1||^2 \leq ||e(Y, X_1) - \beta_2 e(X_2, X_1) - \ldots - \beta_n e(X_n, X_1)||^2.$$

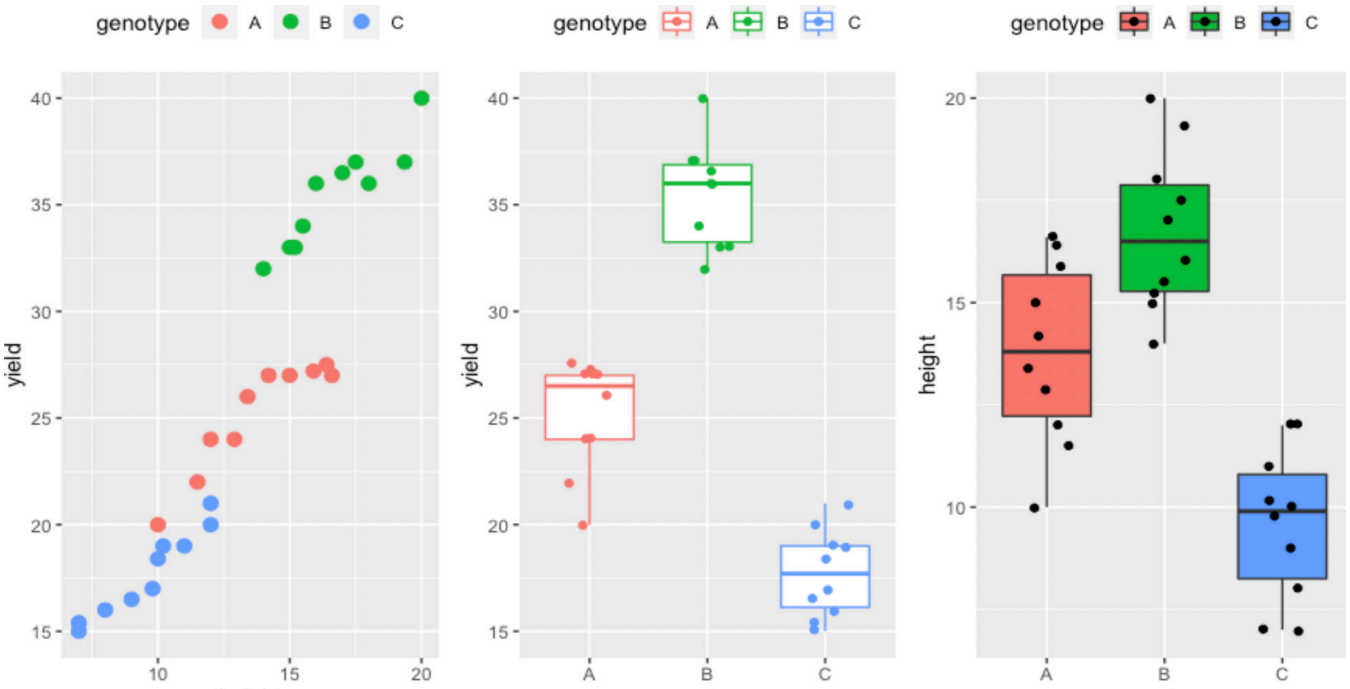


Continuing this pattern, we can further express the residuals in the equations as residuals of residuals, and so on. That is, we predict the value of ($e(Y, X_1)$ - $\beta_2 \cdot e(X2, X_1)$ - ... - $\beta_{i-1} \cdot e(X_{i-1}, X_1)$ - $\beta_{i+1} \cdot e(X_{i-1}, X_1)$ - ... - $\beta_n \cdot e(X_n, X_1)$) using $e(X_i, X_1)$. Without loss of generality, let $i = 2$. So, we obtain:

$$||e(Y, X_1) - \beta_2 e(X_2, X_1) - \ldots - \beta_n e(X_n, X_1)||^2 \leq$$
$$||e(e(Y, X_1), e(X_2, X_1)) - \beta_3 e(e(X_3, X_1), e(X_2, X_1)) - \ldots - \beta_n e(e(X_n, X_1), e(X_2, X_1))||^2,$$

thus also eliminating β_2. At last, we obtain an expression with only two residual terms given by $e(e(...), e(...))$ - $\beta_n \cdot e(e(...), e(...))$ with the first term relating ultimately to the residual of Y with X_1, and the last term relating ultimately to the residual of X_n to X_1. This way, we can calculate the inverse of the design matrix X and directly calculate the optimal value of β by first finding the optimized value of β_n, as every other value of β_i depents on all the values of β_j, for $i \leq j \leq n$. ■

ANCOVA & COVARIANCES IN REGRESSION



ANCOVA *vs.* Linear Regression

ANCOVA is a statistical method used to test for significant differences between means in two or more groups. On the other hand, linear regression is a method used to model the relationship between a dependent variable and one or more independent variables.

ANCOVA and linear regression are related in that they both use the same basic framework of partitioning the total sum of squares into different components. ANCOVA can be thought of as a special case of linear regression where the independent variable is categorical.

Categorical Label in Regression

Many times we encounter cases in which there are underlying confounding variables that will affect the output of our regression model. For instance, in a case where we want to predict the score of students from the number of hours they spent reviewing, and there are three different, independent classes, then the classes that

students attend is a possible confounding variable. We might obtain a graph that appears to be downward sloping, because students in certain classes are given lower scores in general but review for longer hours, but within each group, there is a positive correlation. Thus, it would be better to split our regression model into these multiple groups and analyze their variances and coefficients, respectively.

Let's first consider the scenario with two different groups. Let X be the original design matrix and Y be the column vector that can be divided into two independent groups. Define W as the matrix $[Z\ X]$, where

$$Z = \begin{bmatrix} J_{n_1} & O_{n_1} \\ O_{n_2} & J_{n_2} \end{bmatrix},$$

where O_n be a column vector consisting of n 0s and $n_1 + n_2 = n$, meaning the sum of the length of data in two groups is the full length of the dataset. Let β be the common slope across both groups when we fit Y against X. Define Γ as the following matrix:

$$\Gamma = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \beta \end{bmatrix},$$

where μ_1 is the intercept of group 1, and μ_2 is the intercept of group 2.

If we minimize the new least squares expression $||Y - W\Gamma||^2 = ||Y - X\beta - Z\mu||^2$, where μ is the column vector consisting of μ_1 and μ_2. If we fix the value of $(Y - X\beta)$ and use Z to predict the value of $(Y - X\beta)$. The maximum likelihood estimate for μ is the mean of $(Y - X\beta)$. Let $\bar{Y_1}$ and $\bar{Y_2}$ be the column vectors of the two groups in our label, and $\bar{X_1}$ and $\bar{X_2}$ be the matrix representation of two groups in our design matrix, which are of length n_1 and n_2, respectively. So, $\mu_1 = \bar{Y_1} - \bar{X_1}\beta$ and $\mu_2 = \bar{Y_2} - \bar{X_2}\beta$. If we substitute the above results back to our least squares equation, we get:

$$||Y - Z\Gamma - X\beta||^2 = || \begin{bmatrix} Y_1 - \bar{Y_1} \cdot J_{n_1} \\ Y_2 - \bar{Y_2} \cdot J_{n_2} \end{bmatrix} - \begin{bmatrix} X_1 - \bar{X_1} \\ X_2 - \bar{X_2} \end{bmatrix} \beta||^2,$$

from which we can derive the maximum likelihood estimation of β with the design matrix and output vector both divided into two groups.

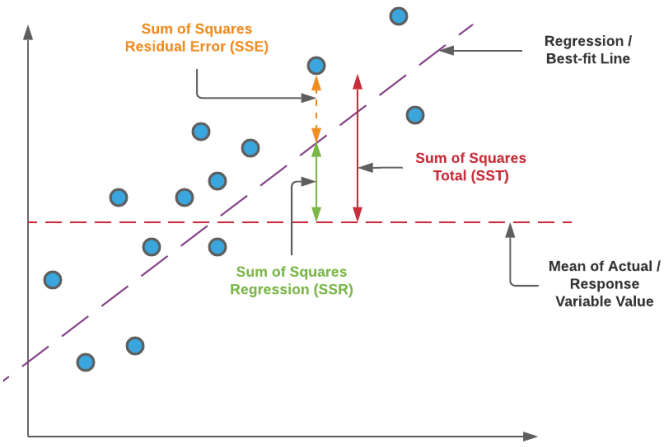
The MLE of β follows

$$\hat{\beta} = \frac{<Y - \bar{Y}, X - \bar{X}>}{<X - \bar{X}, X - \bar{X}>} = \frac{\sum_{i=1}^{2} \sum_{j=1}^{n_i} (y_{ij} - \bar{y_j})(x_{ij} - \bar{x_j})}{\sum_{i=1}^{2} \sum_{j=1}^{n_i} (x_{ij} - \bar{x_j})^2}$$

for each column of X. We observe that the numerator of the expression contains the summation of the MLE of the coefficient matrix when we fit X_1 with Y_1 and when we fit X_2 with Y_2. We call these coefficient matrices β_1 and β_2, respectively. Henceforth, $\beta = P\beta_1 + (1\text{-}P)\beta_2$, where $P = (\Sigma_j\ (x_{1j} - \bar{x_1}))/(\Sigma_i \Sigma_j\ (x_{ij} - \bar{x_i})^2)$, since the denominator can be seen as a weighted average between the two groups. Thus, we have derived the MLE for β that adjusted for the baseline weight/coefficient for the two groups. ■

INTUITION OF THE CORRELATION COEFFICIENT

SSR, SSE, and SST



The Sum of Squares Regression (SSR) is defined as the summation of the squares of the difference between the fitted y values and the mean of observed y values, which is a measure of the amount of variation in y that is explained by the linear model, which follows $\Sigma(\bar{y}\text{-}\hat{y})^2$.

The Sum of Squared Errors/Residuals (SSE) is the summation of all the squared residuals, which follows $\Sigma(y\text{-}\hat{y})^2$.

The Total Sum of Squares (SST) is defined as a statistical measure of deviation from the mean, which follows $\Sigma(y\text{-}\bar{y})^2$.

Proof of "SST = SSE + SSR" & Partition of Variances

We can write the design matrix X with a $[J_n, X']$, with J_n being a column matrix composed of n 1s and X' the rest of the X matrix.

The SSR of the model can be expressed as:

$$
\begin{aligned}
&||\bar{Y} - \hat{Y}||^2 \\
&= ||H_J \bullet Y - H_X \bullet Y||^2 \\
&= ||(H_J - H_X)Y||^2 \\
&= Y^T(H_J - H_X)^T(H_J - H_X)Y \\
&= Y^T(H_J - H_X)(H_J - H_X)Y \\
&= Y^T(H_J^T H_J + H_X^T H_X - H_J^T H_X - H_X^T H_J)Y \\
&= Y^T(H_J + H_X - 2H_J H_X)Y,
\end{aligned}
$$

as the hat matrices have the property that $H^T H = HH = H$. From the fact that the dot product of the residuals and any vector lying on a linear combination of the design matrix X is equal to zero, as proven by the fact that for a linear transformation of the X ($X\beta = \Gamma$), $e \bullet X\beta = Y(I\text{-}H_X)X\beta = Y(X\text{-}H_X X)\beta = Y(X\text{-}X)\beta$ as $H_X X = X$, and that J_n is always proportional to the coefficient column in X, so it lies on some $X\beta$, we obtain that $e \bullet J_n = Y(I\text{-}H_X)J_n = 0$. This further implies $(I\text{-}H_X)J_n = 0$ as $Y \neq 0$. We then obtain:

$$J_n = H_X J_n$$

$$J_n(J_n^T J_n)^{-1} J_n^T = H_X J_n (J_n^T J_n)^{-1} J_n$$
$$H_J = H_X H_J$$
$$H_J = H_J H_X.$$

Hence, the expression for SSR $Y^T(H_J + H_X - 2H_J H_X)Y$ further evaluates to

$$Y^T(H_J H_X + H_X - 2H_J H_X)Y = Y^T(H_X - H_J H_X)Y = Y^T(H_X - H_J)Y.$$

The SSE can be expressed as:

$$\begin{aligned} & ||Y - \hat{Y}||^2 \\ &= ||Y - X(X^T X)^{-1} X^T Y||^2 \\ &= ||(I - X(X^T X)^{-1} X^T)Y||^2 \\ &= Y^T(I - H_X)Y. \end{aligned}$$

The SST can be expressed as:

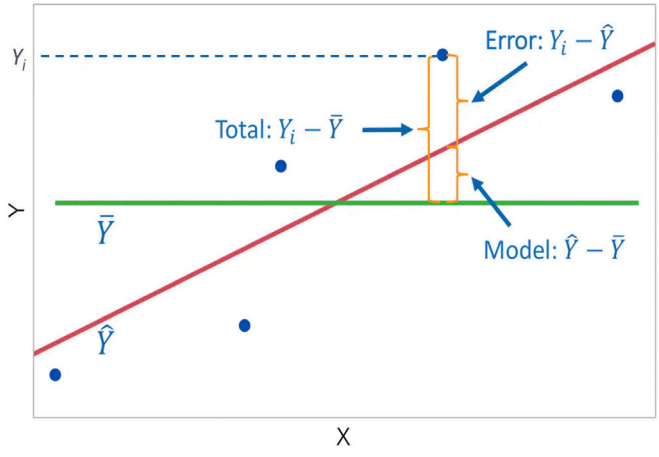
$$\begin{aligned} & ||Y - \bar{Y}||^2 \\ &= ||Y - J_n(J_n^T J_n)^{-1} J_n^T Y||^2 \\ &= ||Y - H_J \bullet Y||^2 \\ &= ||(1 - H_J)Y||^2 \\ &= Y^T(I - H_J)^T(I - H_J)Y \\ &= Y^T(I - H_J)Y, \end{aligned}$$

as $(I-H_J)$ is idempotent. The SST can further be expressed as the summation of SSR and SSE:

$$\begin{aligned} & ||Y - \bar{Y}||^2 \\ &= Y^T(I - H_J)Y \\ &= Y^T(I - H_X + H_X - H_J)Y \\ &= Y^T(I - H_X)Y + Y^T(H_X - H_J)Y \\ &= \text{SSE} + \text{SSR}. \end{aligned}$$

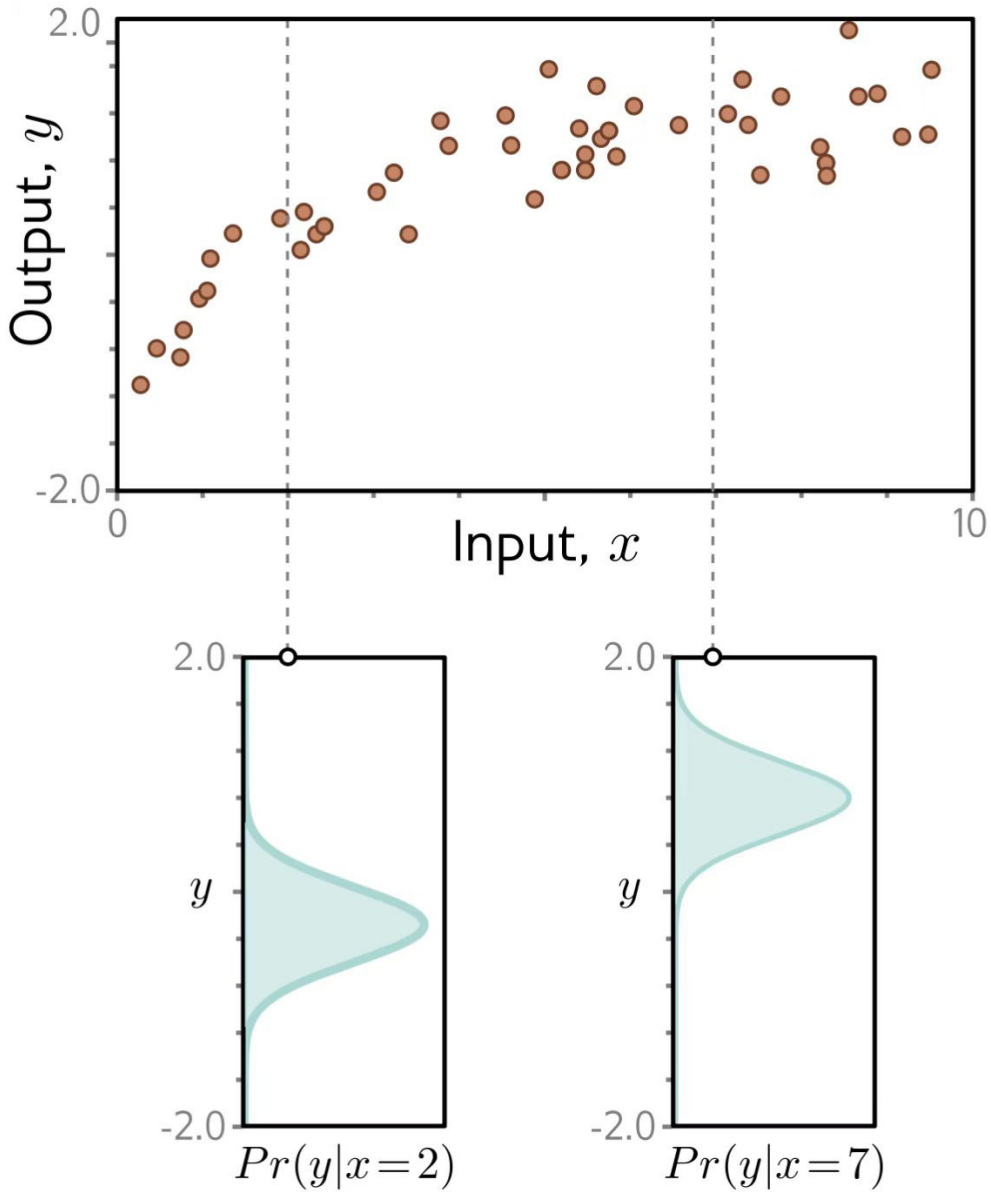
Therefore, we have completed the proof that SST = SSE + SSR.

From the result above, we can partition the variance of our linear model into the variance of the residuals and the variance of the regression model. The square of the correlation coefficient, r^2, is defined as the quotient of SSR and SST (SSR/SST). It can be interpreted as the proportion of total variability explained by the linear association with the added regressors.



Intuitively, it also follows that the larger the SSR, the more variability in the set of data can be attributed to the model, and thus the greater the value of r^2. ■

DISTRIBUTION OF THE PREDICTED Ŷ



Chi-Squares In Quadratic Forms

We know that for an $X \sim N(\mu, \Sigma)$, where Σ is the variance-covariance matrix. Then, $(X-\mu)\Sigma^{-1/2} \sim N(0, 1)$, a standard normal distribution. The square of each i.i.d. data point in $(X-\mu)\Sigma^{-1/2}$ follows a chi-squared distribution; that is, $(X-\mu)^T\Sigma^{-1}(X-\mu) \sim \chi^2$ with n degrees of freedom.

Interested in learning more about this magazine?
Contact the author at sophia.yx.zhu@gmail.com.

Random variables following a ***normal/Gaussian distribution*** can be expressed as $\frac{X-\mu}{\sigma}$, where σ is the population standard deviation.
Random variables following a ***Student's t-distribution*** can be expressed as $\frac{X-\mu}{S}$, where S is the sample standard deviation.
Random variables following a ***chi-squared distribution*** can be expressed as $\frac{S^2}{\sigma^2}$.
A t-distribution can be seen as a normal distribution over the square root of a χ^2 distribution divided by its degrees of freedom.

In fact, if we have a matrix A such that $A\Sigma$ is idempotent, then $(X-\mu)^T A(X-\mu)$ also follows a chi-squared distribution, as proved by the following.

Let p be the rank of A and n be the rank of X with $p < n$. Since A is idempotent, we have $A\Sigma A\Sigma = A\Sigma$, which further implies $A\Sigma A = A$.

If we write out the eigenvalue decomposed version of $A = VDV^T$, where V is of size $n \times p$ and D is of size $p \times p$, we can rewrite $A\Sigma A$ as $VDV^T\Sigma VDV^T = A = VDV^T$. This equation further implies:

$$V^TVDV^T\Sigma VDV^TV = V^TVDV^TV.$$

By the property that $V^TV = I$ (as V is an orthogonal matrix of eigenvectors), we can simplify the above equation:

$$DV^T\Sigma VD = D$$
$$D^{1/2}V^T\Sigma VD^{1/2} = I.$$

We then create the random variable defined as $D^{1/2}V^T(X-\mu)$, which follows $N(0, D^{1/2}V^T\Sigma VD^{1/2}) = N(0, 1)$. The square of this i.i.d. normal distribution $(X-\mu)^TV^TD^{1/2}D^{1/2}V^T(X-\mu) = (X-\mu)^TA(X-\mu) \sim \chi^2$ with p degrees of freedom.

Variance of Residuals & Chi-Squares

In this section, we will show that the dot product of the residuals e to itself divided by the population variance follows a chi-squared distribution using the previous result.

$$\frac{e^Te}{\sigma^2} = \frac{Y^T(I-H_X)Y}{\sigma^2} = \frac{(Y-X\beta)^T(I-H_X)(Y-X\beta)}{\sigma^2}.$$

The second part of the equation is valid as $HxX = 0$, so the additional $X\beta$ term does not add anything to the previous equation. The expression is of the form $(Y-\bar{Y})^TA(Y-\bar{Y})$ where $A = (I-Hx)/\sigma^2$. Check that $A\Sigma = (I-Hx)/\sigma^2 \times (\sigma^2 I) = (I-Hx)$, an idempotent matrix. Thus, $\frac{e^Te}{\sigma^2} \sim \chi^2$. Another way to write this expression is $\frac{(n-p)S^2}{\sigma^2}$. Its degrees of freedom is $\text{rank}(I-Hx) = n-p$ for a design matrix X with a rank of p.

T-Distributed Coefficients

Let q be a column vector that composes of all zeroes except for certain positions that we want to observe in β and β be the coefficient matrix. For instance, if we want to only pick out the i^{th} column in β, then we will only assign the i^{th} position in q to be 1. The covariance of $q^T\beta$ and the residuals e is $\text{cov}(q^T\beta, e) = \text{cov}(q^T(X^TX)^{-1}X^TY, (I-Hx)Y) = q^T(X^TX)^{-1}X^T \cdot \text{cov}(Y, Y) \cdot (I-Hx)^T = q^T(X^TX)^{-1}X^T(I-Hx) \cdot \sigma^2$. Since $X^T(I-Hx) = 0$, the residuals are orthogonal (independent) to $q^T\beta$.

We know that the $q^T\beta \sim N(q^T\beta, q^T(X^TX)^{-1}q\,\sigma^2)$, since we assume normality in our residuals and thus normality in our response variable Y, as well as the fact that the least squares equation is based on the MLE of normal curves. $\frac{q^T\hat{\beta}-q^T\beta}{\sigma^2} \sim N(0,1)$. If we divide the standard normal above by the square root of the chi-squared distribution of $\frac{(n-p)S^2}{\sigma^2}$ divided by the degrees of freedom, we get a t-distribution as the following:

$$\frac{\frac{q^T\hat{\beta}-q^T\beta}{\sigma^2} \sim N(0,1)}{\sqrt{\frac{(n-p)S^2}{(n-p)\sigma^2}} \sim \sqrt{\frac{\chi^2_{n-p}}{df=n-p}}} = \frac{q^T\hat{\beta}-q^T\beta}{S\sqrt{q^T(X^TX)^{-1}q}} \sim t_{n-p},$$

showing that each coefficient in β follows a t-distribution if we create an interval using the sample standard deviation. ■

DERIVATION OF THE F-STATISTIC AND TEST

Why F-Test?

If we want to compare whether a column in X is significantly correlated with Y, we consider using the F-statistics defined as:

$$\frac{(K\hat{\beta}-K\beta)^T(K(X^TX)^{-1}K^T)^{-1}(K\hat{\beta}-K\beta)}{\operatorname{rank}(K)S^2},$$

where K is a full row rank contrast matrix that tests whether a linear combination of certain columns in X is a significant predictor of Y (or whether the coefficients for a group of predictor variables are jointly equal to zero).

The F-statistic is calculated by comparing the variance of the residuals in the reduced model (i.e., the model without the predictor(s) associated with the null hypothesis) to the variance of the residuals in the full model (i.e., the model with all predictors). The F-statistic is the quotient of the two chi-squared distributions, with the numerator being the distribution of the variance of the reduced model and the denominator being the distribution of the variance of the full model, weighted by the degrees of freedom associated with each model. According to the formula above, the F-statistic can also be seen as the square of a t-distribution divided by rank(K), or the degrees of freedom of the reduced model.

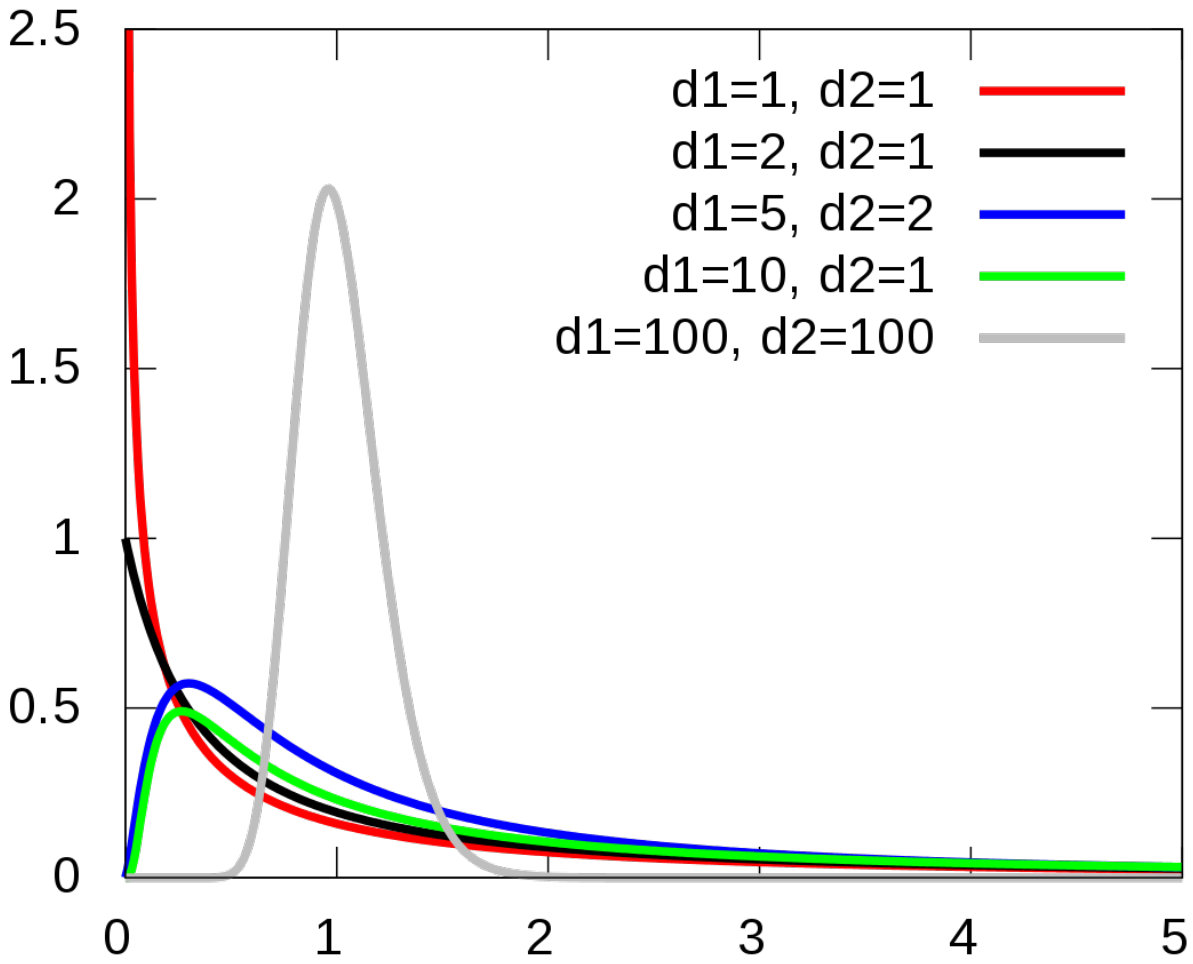
What F-Test?

The intuition behind F is the ratio of the variances of two models. The test follows H_0: $\beta_1 = \beta_2 = ... = \beta_n = 0$, or $K\beta = 0$, against H_A: at least 1 $\beta \neq 0$., or $K\beta \neq 0$. If our model is a reduced model with p number of predictors, we want to determine if the reduced model, a subset of the predictor variables in a full model, significantly contributes to the prediction of the response variable compared to the full model.. The standard F-statistics is:

$$\frac{MSR}{MSE},$$

in which the numerator is the SSR divided by the degrees of freedom of the SSR of the reduced model (p) and the denominator is the SSE divided by the degrees of freedom of the SSE of the reduced model (n-p), in which n is the total number of observations. Both the numerator and the denominator are chi-squared due to the residual normality assumption divided by their degrees of freedom, a scaling factor that makes the quotient justifiable. Since MSR measures the variation in the response variable explained by the regression model, and MSE measures the unexplained variation in the response variable, the statistic is a signal-to-noise ratio.

The SSR of the reduced model is:



$(K\hat{\beta}-K\beta)^T(K(X^TX)^{-1}K^T\sigma^2)^{-1}(K\hat{\beta}-K\beta)$, which follows a chi-squared distribution with rank(K) = p degrees of freedom. The numerator of the F-statistics, or the mean squares of regression, is:

$$\frac{(K\hat{\beta}-K\beta)^T(K(X^TX)^{-1}K^T\sigma^2)^{-1}(K\hat{\beta}-K\beta)}{\text{rank}(K)}$$

.

The denominator, on the other hand, is the sum of squared residuals divided by its degrees of freedom (n-p). It is the distribution $\frac{(n-p)S^2}{\sigma^2}$ further divided by (n-p) = (n - rank(K)). As a result, the F-statistic for testing whether $K\beta = 0$ simplified to:

$$\frac{(K\hat{\beta}-K\beta)^T(K(X^TX)^{-1}K^T)^{-1}(K\hat{\beta}-K\beta)}{\text{rank}(K)S^2}$$

.

Try This Question:

Which distribution does the expression $0.5[(K\hat{\beta}-m)^T(K(X^TX)^{-1}K^TS^2)^{-1}(K\hat{\beta}-m)]$, follow if we are testing against the null hypothesis that $K\beta = m$?

1. A χ^2 distribution with n-rank(K) degrees of freedom.
2. The square of a t-distribution with n-rank(K) degrees of freedom divided by 2.
3. A F-distribution with 2 numerator degrees of freedom and n-rank(K) denominator degrees of freedom. ■

■ ***Contact the author to learn more.***

PREDICTION INTERVALS

Interested in learning more about this magazine?
Contact the author at sophia.yx.zhu@gmail.com.

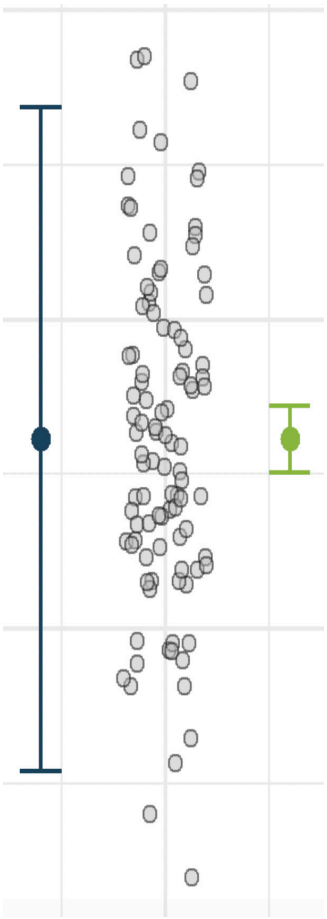
Confidence *vs.* Prediction Intervals

A prediction interval is a statistical measure that estimates an interval in which future observations are likely to fall based on a given set of predictors. It is similar to a confidence interval, but a confidence interval estimates the range of values that contains a population parameter with a certain level of confidence, a prediction interval estimates the range of values that contains an individual observation with a certain level of confidence. While the 95% confidence interval of $\hat{y}$ is $\hat{y}_0 \pm t_{95\%,\, df = n-p} \cdot S(x_0^T(x_0^T x_0)^{-1}x_0)^{1/2}$, where the variance is Var($\hat{y}_0$), the variance of prediction intervals for regression coefficients are derived as follows:

$$\begin{aligned} &\mathrm{Var}(e) \\ &= \mathrm{Var}(y - \hat{y}_0) \\ &= \mathrm{Var}(y - x_0\beta) \\ &= \sigma^2 + (x_0^T(x_0^T x_0)^{-1}x_0)\sigma^2 \\ &= (1+x_0^T(x_0^T x_0)^{-1}x_0)\sigma^2. \end{aligned}$$

Thus, the 95% prediction interval is $\hat{y}_0 \pm t_{95\%,\, df = n-p} \cdot S(1+x_0^T(x_0^T x_0)^{-1}x_0)^{1/2}$, where Var($e$) is used since Var($\hat{y}_0$) can be separated into two parts: the variance explained by the prediction model and the unexplained error variance Var(e), and we only want to include Var(e) that captures the variation of the individual response around the predicted response.

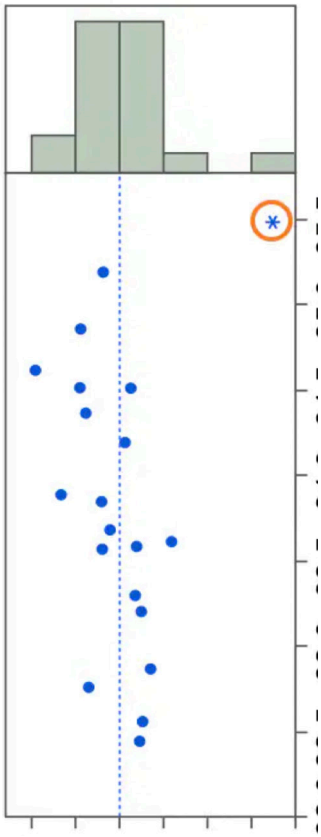
Prediction intervals don't converge to a single point estimate like confidence intervals. In other words, a confidence interval accounts for the uncertainty in estimating a population parameter from the sample, while a prediction interval accounts for the additional uncertainty in predicting the value of a future observation from the population. This additional uncertainty arises from the inherent variability of the response variable, which cannot be fully captured by the sample data alone. ■



Confidence Interval

Prediction Interval

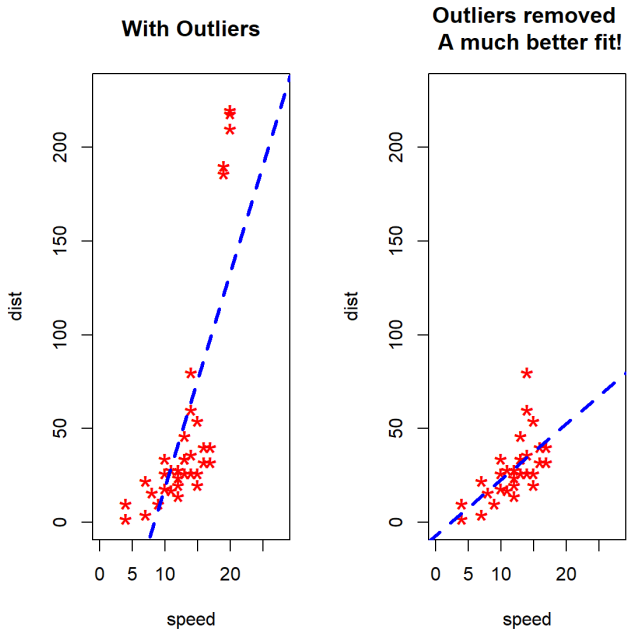
LEAVE-ONE-OUT RESIDUALS



Outliers & Leave-One-Out Residuals

Usually, we want to eliminate the noises (data points that are contaminated or provide false influence on our linear models). These outliers are usually found using the 1.5 × IQR rule or the within 2 standard deviation rule (within 95% of the spread assuming normality).

Leave-One-Out (LOO) residuals are a type of cross-validation method used in regression analysis to evaluate the performance of a model. The LOO method involves removing a single observation from the dataset, fitting the model on the remaining n-1 observations, and using the model to predict the response variable for the left-out observation. LOO residuals can be used to identify potential outliers or influential observations. An observation with a large LOO residual indicates that the model fit is highly sensitive to that observation, and its removal could result in a substantially different model. Such an observation is considered influential and could be an outlier. A more common usage of LOO is to assess the performance of a model by calculating metrics such as the Root Mean Squared Error (RMSE) or the Mean Absolute Error (MAE) based on the n LOO residuals. These metrics provide a measure of how well the model is able to predict the response variable for new observations not included in the original dataset.



The LOO residual is calculated as $y_i - \hat{y}_i^{(-i)}$, where $\hat{y}_i^{(-i)}$ is the predicted value for observation i from the model fitted on the remaining n-1 observations. Let δ_i be a column vector with all 0s except for an 1 on the i^{th} position. W = $[X\ \delta_i]$, and

$$\Gamma = \begin{bmatrix} \beta \\ \Delta_i \end{bmatrix},$$

where Δ_i is the coefficient appended to the coefficient matrix β for the additional δ_i column in W. Then, the nuance behind LOO in matrix form is presented below.

The least squares equation now becomes $||Y\text{-}W\Gamma||^2 = \Sigma_j (y_j - \Sigma_k x_{j,k}\beta_k - \delta_{i,j}\Delta_i)^2 = \Sigma_{j\neq i} (y_j - \Sigma_k x_{j,k}\beta_k)^2 + (y_i - \Sigma_k x_{i,k}\beta_k - \Delta_i)^2$. By separating the loss function into two parts, we can minimize the two parts respectively. The MLE of Δ_i is $y_i - \Sigma_k x_{i,k}\beta_k$ if we let $(y_i - \Sigma_k x_{i,k}\beta_k - \Delta_i)^2 = 0$. So, we get $\Sigma_{j\neq i} (y_j - \Sigma_k x_{j,k}\beta_k)^2 + (y_i - \Sigma_k x_{i,k}\beta_k - \Delta_i)^2 \geq \Sigma_{j\neq i} (y_j - \Sigma_k x_{j,k}\beta_k)^2 = ||Y^{(-i)} - X^{(-i)}\beta||^2$. The MLE of β in the new least squares equation is $\hat{\beta}^{(-i)}$. Thus, $||Y^{(-i)} - X^{(-i)}\beta||^2 \leq ||Y^{(-i)} - X^{(-i)}\hat{\beta}^{(-i)}||^2$. Returning to the MLE of Δ_i, we have:

$$\hat{\Delta}_i = y_i - \sum_k x_{i,k}\hat{\beta}_k^{(-i)} = y_i - \hat{y}_i^{(-i)}.$$

This result shows that:

1. Adding a regressor that is all 0 but an 1 in the ith data point is equivalent to deleting that data point from analysis. So, if the i^{th} data point is a potential outlier or noise, we can simply append another column δ_i into our design matrix X.
2. The coefficient for δ_i is equivalent to the LOO residual of the i^{th} data point. The LOO residual can also be seen as the magnitude of how different the observed data point is from the fitted model without the i^{th} data point, or how much it deviates from the average trend of the dataset.

PRESS Residuals & Error Distribution

PRESS (Predicted Residual Sum of Squares) residuals are a measure of the predictive power of a linear model. They are computed by omitting one observation at a time from the data set, and then calculating the predicted value for the omitted observation. The PRESS residual is the difference between the actual value of the omitted observation and its predicted value based on the reduced model.

Let's start by rewriting the design matrix X as rows:

$$X = \begin{bmatrix} Z_1 \\ \vdots \\ Z_n \end{bmatrix}.$$

The quantity $X^TX = \Sigma_j z_j z_j^T$, while $X^{(-i)T}X^{(-i)} = \Sigma_{j\neq i} z_j z_j^T$. Therefore, $X^{(-i)T}X^{(-i)} = X^TX - z_i z_i^T$. The inverse of $X^{(-i)T}X^{(-i)}$ can be found using the Sherman-Morrison-Woodbury Theorem as:

$$(X^{(-i)T}X^{(-i)})^{-1} = (X^TX)^{-1} + \frac{(X^TX)^{-1}Z_iZ_i^T(X^TX)^{-1}}{1-Z_i^T(X^TX)^{-1}Z_i}.$$

Since $H_X = X(X^TX)^{-1}X^T$, the i^{th} diagonal slot $H_{X\,i,i} = \delta_i^T X(X^TX)^{-1}X^T\delta_i = z_i^T(X^TX)^{-1}z_i$. Substitute the value $H_{Xi,i}$ into the expression for $(X^{(-i)T}X^{(-i)})^{-1}$ to get:

$$(X^{(-i)T}X^{(-i)})^{-1} = (X^TX)^{-1} + \frac{(X^TX)^{-1}Z_iZ_i^T(X^TX)^{-1}}{1-H_{Xi,i}}.$$

The quantity $X^{(-i)T}Y^{(-i)} = X^TY - z_iy_i$ by the same logic used to deduce $X^{(-i)T}X^{(-i)}$. We also know that $\hat{y}_i^{(-i)} = z_i^T(X^TX)^{-1}X^{(-i)T}Y^{(-i)}$, with z_i^T being the i^{th} row in X. As we know, the value of $(X^{(-i)T}X^{(-i)})^{-1}$ and $X^{(-i)T}Y^{(-i)}$, $\hat{y}_i^{(-i)}$ can be written as:

$$Z_i^T((X^TX)^{-1} + \frac{(X^TX)^{-1}Z_iZ_i^T(X^TX)^{-1}}{1-H_{Xi,i}})(X^TY - Z_iy_i)$$
$$= \hat{y}_i + \frac{H_{Xi,i}}{1-H_{Xi,i}} - H_{Xi,i}y_i - \frac{H_{Xi,i}^2}{1-H_{Xi,i}}y_i,$$

which after expansion equates to $\frac{\hat{y}_i}{1-H_{Xi,i}} + y_i + \frac{y_i}{1-H_{Xi,i}}$. This implies $y_i - \hat{y}_i^{(-i)} = \frac{e_i}{1-H_{Xi,i}}$ and that LOO residuals can be computed without refitting the model and follows the distribution of the i^{th} error multiplied by the sample variance. ■

CONFIDENCE ELLIPSOIDS

Ellipsoid In Vector Form

Any arbitrarily oriented ellipsoid can be written as $(X\text{-}v)^T A(X\text{-}v) = 1$, with the center at v.

Relation With F

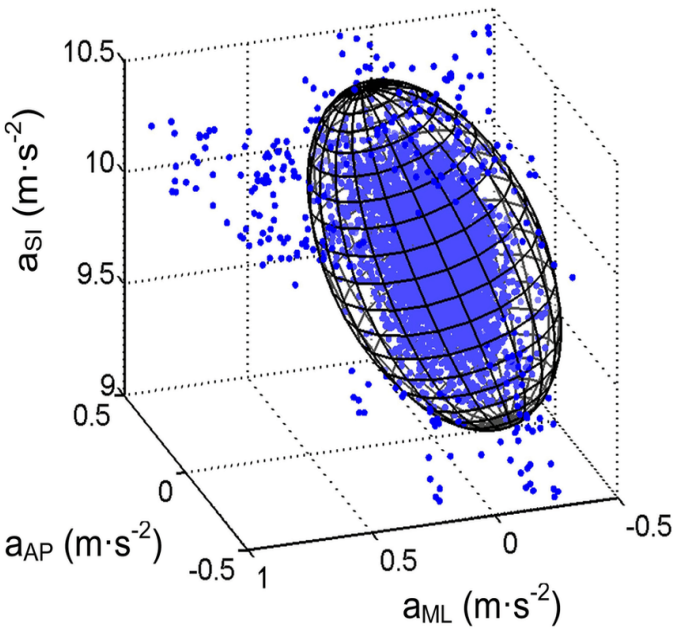
Continuing our discussion on F-tests, if we want to test H_0: $K\beta = 0$ against H_A: $K\beta \neq 0$, it is equivalent to deciding whether our F-statistic

$$\frac{(K\hat{\beta}-K\beta)^T(K(X^TX)^{-1}K^T)^{-1}(K\hat{\beta}-K\beta)}{\text{rank}(K)S^2}$$

is within a certain range of the F-distribution (e.g., 95% confidence interval). Let $K\beta = m$, and rank$(K) = v$, the F-statistic becomes

$$\frac{(K\hat{\beta}-m)^T(K(X^TX)^{-1}K^T)^{-1}(K\hat{\beta}-m)}{vS^2}.$$

Bounding the statistic with the critical F-values forms an ellipsoid: $\frac{(K\hat{\beta}-m)^T(K(X^TX)^{-1}K^T)^{-1}(K\hat{\beta}-m)}{vS^2} \leq F_{95\%}(K(X^TX)^{-1}K^T, S^2)$, which deduces $\frac{(K\hat{\beta}-m)^T(K(X^TX)^{-1}K^T)^{-1}(K\hat{\beta}-m)}{vS^2 \cdot F_{95\%}(K(X^TX)^{-1}K^T,S^2)} \leq 1$. With A being $\frac{(K(X^TX)^{-1}K^T)^{-1}}{vS^2 \cdot F_{95\%}(K(X^TX)^{-1}K^T,S^2)}$, the result forms the region bounded inside an ellipsoid. Thus, we obtain a "confidence ellipsoid," in which are the values of m that we fail to reject our null hypothesis. In other words, m is in a certain range around v such that there is not enough evidence to reject H_0.



Summary

A confidence ellipsoid is a geometric representation of the confidence intervals of multiple regression coefficients. It is an ellipsoid whose center corresponds to the estimated values of the coefficients, and whose shape and size are determined by the standard errors of the estimates and the level of confidence desired. The ellipsoid represents the set of values that the coefficients can take with a given level of confidence, and can be used to test hypotheses and make predictions about the values of the coefficients. A smaller ellipsoid indicates higher precision of the estimates, while a larger ellipsoid indicates greater uncertainty. ■

FREQUENTIST STATISTICS

MONTE CARLO SIMULATION & PROBABILITY

The Jackknife

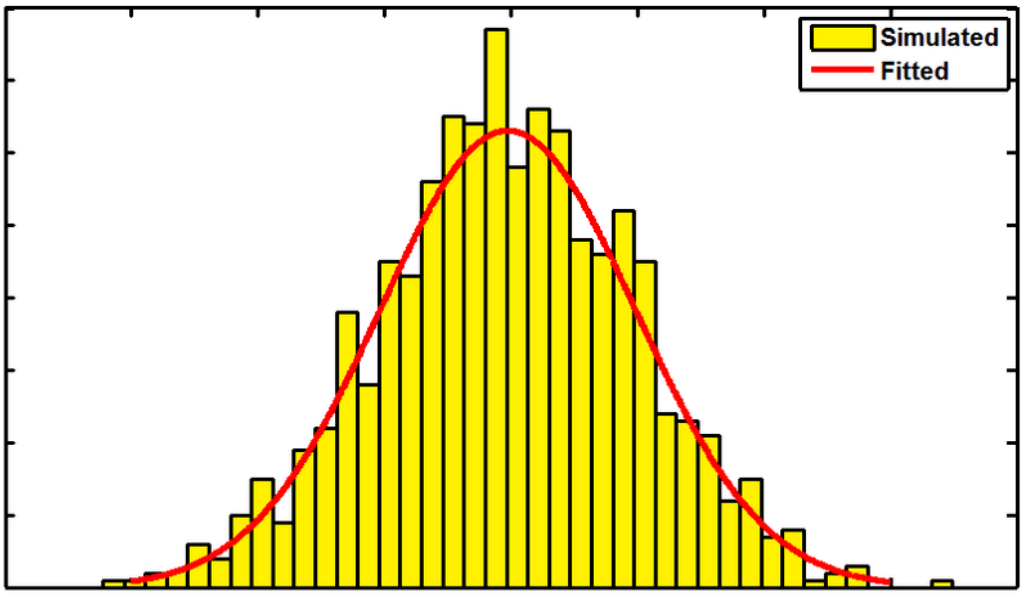
Bootstrapping is a statistical resampling technique used to estimate the distribution of a statistic. It involves repeatedly leaving one observation out of the sample, computing the statistic of interest for each subsample, and then combining the results to estimate the population statistic. The statistic of interest can be calculated from these resampled datasets, and the distribution of the statistic across the resampled datasets is used to estimate its sampling distribution.

Monte Carlo Simulations

Monte Carlo simulation is a computational technique that uses random sampling to simulate complex systems or processes. The method is named after the famous Monte Carlo Casino, which is known for its games of chance and randomness. The Jackknife can be considered a specific type of Monte Carlo simulation. Monte Carlo is useful in situations where the underlying population distribution is unknown or difficult to model, or where the sample size is small and traditional methods of inference may not be reliable.

Try These Questions:

1. Estimate the value of π using the normal approximation of the binomial distribution by sampling the proportion points that lies within a quarter of a circle.
2. Estimate the value of π using the density function $\sqrt{1-X^2}$ and computing the area. Which method yields smaller variance? ■



The Monte Carlo Casino

INTUITION BEHIND THE POISSON DISTRIBUTION

Poisson Modeling

The Poisson distribution is a probability distribution that is often used to model the number of times an event occurs in a fixed interval of time or space, given the average rate at which the event occurs. It is named after the French mathematician Simeon Denis Poisson, who introduced it in the early 19th century.

The intuition behind the Poisson distribution can be understood with an example. Consider a factory that produces light bulbs. The factory produces light bulbs at a rate of 10 bulbs per minute. We want to know the probability of a certain number of bulbs being produced in a given time interval, say 1 minute.

The Poisson distribution follows:

$$P(X = k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!},$$

where X is the random variable representing the number of events, k is the number of events, and λ is the average rate at which the events occur.

Mathematical Derivation

The Poisson distribution is to resolve the limitations of a binomial distribution, which can only model binary data.

The binomial PDF follows:

$$binom(X = k) = \binom{n}{k} p^k (1-p)^{1-k}.$$

The main difference directly observed from the equation is that the binomial PDF depends on the total number of observations n, while the Poisson PDF does not. This is because we assume the total number of observations tends to infinity in Poisson.

From the definition of p and λ, we have $p = \lambda / n$, as the probability of an event occurring is the quotient of the number of successful events and the total number of events. Then, as n approaches infinity, we must also assume p approaches 0 on an infinitely small interval of time. Thus, the only parameter in the Poisson distribution is λ.

Therefore, we get the following derivation:

$$\begin{aligned}
&P(X = k) \\
&= \lim_{n\to\infty} \binom{n}{k} p^k (1-p)^{n-k} \\
&= \lim_{n\to\infty} \binom{n}{k} (\tfrac{\lambda}{n})^k (1-\tfrac{\lambda}{n})^{n-k} \\
&= \lim_{n\to\infty} \tfrac{n!}{(n-k)!k!} (\tfrac{\lambda}{n})^k (1-\tfrac{\lambda}{n})^{n} (1-\tfrac{\lambda}{n})^{-k} \\
&= \lim_{n\to\infty} \tfrac{n!}{(n-k)!k!} (\tfrac{\lambda}{n})^k (e^{-\lambda}) \cdot 1 \\
&= \lim_{n\to\infty} (\tfrac{n!}{(n-k)!}) (\tfrac{1}{n^k}) (\tfrac{\lambda^k}{k!}) (e^{-\lambda}) \\
&= \lim_{n\to\infty} (\tfrac{n}{n} \tfrac{n-1}{n} \tfrac{n-2}{n} \dots \tfrac{n-k+1}{n}) (\tfrac{\lambda^k}{k!}) (e^{-\lambda}) \\
&= \lim_{n\to\infty} 1 \cdot (\tfrac{\lambda^k}{k!}) (e^{-\lambda}) \\
&= \tfrac{\lambda^k e^{-\lambda}}{k!},
\end{aligned}$$

obtaining the Poisson distribution, which is used in a variety of fields to model the occurrence of rare events or the distribution of counts. ■

RANK SUM & NON-PARAMETRIC HYPOTHESIS TESTS

Tests For Ranks

Rank sum tests are non-parametric tests used to compare two independent samples. The most commonly used rank sum test is the Mann-Whitney U test.

The rank sum test is often used as an alternative to the two-sample t-test when the data are not normally distributed or when the assumptions of the parametric tests are violated, such as small sample sizes and unequal variances.

Mann-Whitney U Test

The Mann-Whitney U test evaluates whether two samples have the same distribution by comparing the ranks of the data in each sample. The null hypothesis of the test is that there is no difference between the distributions of the two groups. In other words, we assume that both groups are i.i.d. distributed.

The test works by assigning ranks to all the observations in the two groups and calculating the sum of the ranks for each group. To derive the test statistic, we first calculate the expected value and variance of the ranks. Let n_A and n_B denote the sample sizes of the two groups A and B, and W be the sum of the ranks of group A. Then, under the null hypothesis that the ranks of A are identically distributed, $\mathrm{E}(W) = \mathrm{E}(\sum_{i=1}^{n_A+n_B} iC_i)$,

where $C_i = \begin{cases} 1, & \text{if the } i^{th} \text{ smallest value in } A \cup B \text{ is in } A; \\ 0, & \text{otherwise.} \end{cases}$

So, $\mathrm{E}(W) = P(C_i = 1)\frac{(n_A+n_B)(n_A+n_B+1)}{2}$
$= \frac{n_A}{n_A+n_B}\frac{(n_A+n_B)(n_A+n_B+1)}{2} = \frac{n_A(n_A+n_B+1)}{2}$.

The variance of the test is

$$\begin{aligned} & E(W^2) - E(W)^2 \\ &= \frac{n_A}{n_A+n_B}\frac{(n_A+n_B)(n_A+n_B+1)(2n_A+2n_B+1)}{6} - \frac{n_A^2(n_a+n_B+1)^2}{4} \\ &= \frac{2n_A(n_A+n_B)(2n_A+2n_B+1)-3n_A^2(n_a+n_B+1)^2}{12} \\ &= \frac{n_A n_B(n_A+n_B+1)}{12}. \end{aligned}$$

Thus, we can deduce the normally distributed test statistics U as:

$$\begin{aligned} & \frac{W-\mathrm{E}(W)}{\sqrt{Var(w)}} = \frac{W-\frac{n_A(n_A+n_B+1)}{2}}{\sqrt{\frac{n_A n_B(n_A+n_B+1)}{12}}} \\ &= n_A n_B \frac{n_A(n_A+1)}{2} - \sum_{i=1}^{n_A+n_B} iC_i \sim N(0,1). \end{aligned}$$

The test statistics can similarly be deduced on sample B.

Non-Parametric Tests

Non-parametric tests have the following advantages:

1) The loss in power is usually smaller than their parametric counterparts; 2) They do not require normality assumptions; 3) They are more robust when dealing with outliers.

However, they also have several limitations. Usually, non-parametric tests rely on measures such as rank rather than the raw data, so there will certainly be a loss of information. ■

CHEBYSHEV'S INEQUALITY

Interested in learning more? Contact the author at sophia.yx.zhu@gmail.com.

Bounded Probability

Chebyshev's inequality is a theorem in probability theory that provides a bound on the probability that a random variable deviates from its expected value by more than a certain amount. Specifically, for any random variable X with finite mean μ and finite variance σ^2, Chebyshev's inequality states that the probability that X deviates from μ by more than k standard deviations is at most $1/k^2$, for any positive constant k.

Chebyshev's inequality is a useful tool in probability theory and statistics because it provides a general bound on the probability of extreme deviations of a random variable from its mean, without making any assumptions about the distribution of the variable. However, the bound provided by Chebyshev's inequality is generally not very tight, and stronger bounds can often be obtained using more specific properties of the distribution.

Mathematical Derivation

Mathematically, Chebyshev's inequality can be expressed as:

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

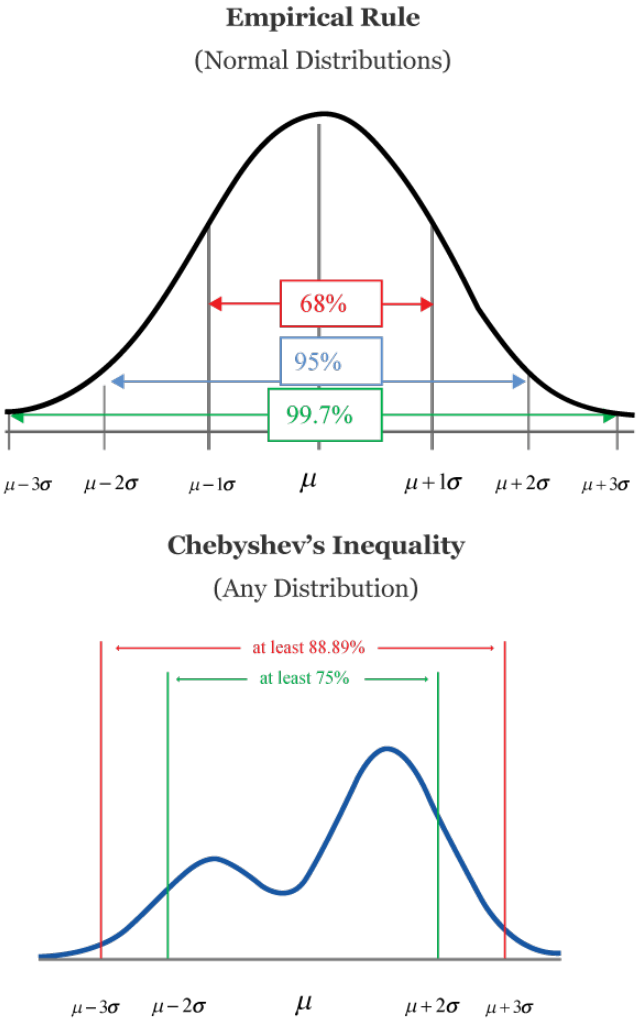
To show the inequality is valid, we can use the following steps:

$$P(|X-\mu| \geq k\sigma) = \int_{X\in\{|X-\mu|\geq k\sigma\}} f(X)dX.$$

Since $|X-\mu| \geq k\sigma \Longrightarrow \frac{|X-\mu|}{k\sigma} \geq 1 \Longrightarrow \frac{(X-\mu)^2}{k^2\sigma^2} \geq 1$, we further derive that:

$$\begin{aligned} P(|X-\mu| \geq k\sigma) &\leq \int_{X\in\{|X-\mu|\geq k\sigma\}} \frac{(X-\mu)^2}{k^2\sigma^2} f(X)dX \\ &\leq \frac{1}{k^2}\int_{-\infty}^{\infty} \frac{(X-\mu)^2}{\sigma^2} f(X)dX \\ &\leq \frac{1}{k^2}\int_{-\infty}^{\infty} Xf(X)dX = \frac{1}{k^2}, \end{aligned}$$

completing the proof. However, the inequality indeed has a relatively loose bound of the probability distribution. ■



THE DELTA METHOD FOR CALCULATING VARIANCE

Estimation of Distribution

The delta method is a statistical technique used to estimate the distribution of a function of a random variable. The method is particularly useful when the random variable is normally distributed or can be approximated by a normal distribution, and the function of the random variable is nonlinear.

The delta method provides an approximation to the distribution of a function of a random variable by using the first and second moments of the random variable. Specifically, let X be a random variable with mean μ and variance σ^2, and let $g(X)$ be a function of X. The delta method states that if X is approximately normal and $g(X)$ is differentiable at μ, then the distribution of $g(X)$ can be approximated by a normal distribution with mean $g(\mu)$ and variance $g'(\mu)^2\sigma^2$, where $g'(\mu)$ is the derivative of $g(X)$ evaluated at μ.

Mathematical Derivation

Mathematically, the delta method can be expressed as that if $\frac{X-\mu}{\sigma} \sim N(0,1)$, $\frac{g(X)-g(\mu)}{g'(\mu)\sigma} \sim N(0,1)$.

The proof is as follows: when X is very close to μ, we get $\frac{g(X)-g(\mu)}{X-\mu} \approx g'(\mu)$ by definition of derivatives. This implies:

$$\frac{g(X)-g(\mu)}{g'(\mu)} \approx X-\mu;$$
$$\frac{g(X)-g(\mu)}{g'(\mu)\sigma} = \frac{X-\mu}{\sigma} \sim N(0,1).$$

Thus, we have completed the proof.

Use Cases

The delta method allows us to approximate the distribution of a function of a random variable using its first-order Taylor series expansion. It is commonly used in statistics to approximate the variance of sample statistics, such as the sample mean or sample proportion, or to estimate the standard error of a regression coefficient. It can also be used to construct confidence intervals and hypothesis tests for parameters in statistical models.

The method is used in many areas of statistics, including econometrics, finance, and engineering. It is particularly useful in estimating the distribution of a nonlinear function of a random variable, such as the standard deviation of a sample mean, the probability of default in a loan portfolio, or the expected return on a stock portfolio. The method provides a simple and computationally efficient way to estimate the distribution of a function of a random variable without resorting to more complex methods such as Monte Carlo simulation. ■

INTUITION BEHIND THE FISHER'S EXACT TEST

Categorical Variables

Fisher's exact test is used to determine the significance of the association between two categorical variables. It is particularly useful when sample sizes are small, and the assumptions of the chi-squared test cannot be met.

Fisher's exact test calculates the probability of observing the observed frequency distribution or a more extreme one. To do this, it uses the hypergeometric distribution, which calculates the probability of observing a certain number of successes (e.g., the frequency of one category of X and one category of Y) in a fixed number of draws (e.g., the total number of observations) from a population with a fixed number of successes (e.g., the total number of observations in the other category of X, n_1, and the other category of Y, n_2).

Hypergeometric Distribution

The probability mass function of the hypergeometric distribution is:

$$P(X = x|X + Y = z) = \frac{\binom{n_1}{x}\cdot\binom{n_2}{z-x}}{\binom{n_1+n_2}{z}}.$$

Specifically, the denominator represents the total number of ways to choose z objects out of all ($n_1 + n_2$) objects, and the numerator represents the joint probability of choosing x objects out of X and choosing the rest z-x objects out of Y.

P-Value of Fisher's Test

Fisher's exact test uses contingency tables to calculate its P-value. For an one-sided fisher's test, the null hypothesis is $p_1 = p_2$, with the alternative hypothesis $p_1 > p_2$ (or the other way around).

Express the data as a contingency table as below, where the first row is X, the second row is Y, the first column is successes, and the second column is failures:

a	b	a+b
c	d	c+d
a+c	b+d	

Then, we write out several other contingency tables with an increasing number of successes in X and a decreasing number of successes in Y, fixing the margins of the contingency tables:

a+1	b-1	a+b		a+b	0	a+b
c-1	d+1	c+d	...	c-b	d+b	c+d
a+c	b+d			a+c	b+d	

Lastly, we calculate the probability of each contingency table occurring using the hypergeometric statistic and sum them up. The sum, which is the P-value of the test, is $\sum_{i=a+1}^{a+b} \frac{\binom{a+b}{i}\binom{c+d}{a+c-i}}{\binom{a+b+c+d}{i}}$. Fisher's exact test gives us an exact P-value and maintains the nominal level of significance without requiring a large sample size. ■

WEIGHTED CONFOUN--DING & THE CMH TEST

Weighting

Simpson's paradox refers to the direction of the association between two variables is reversed when the data is aggregated or grouped by a third variable. Thus, we usually stratify the confounding variable and combine the stratified estimates (like we previously did on ANCOVA of regression).

Suppose we have two normally distributed variables $X \sim N(\mu, \sigma_1^2)$ and $Y \sim N(\mu, \sigma_2^2)$, stratified from the raw data centered at μ. The MLE of μ can be found using the log-likelihood of the joint distribution of X and Y:

$$-(x-\mu)^2/(2\sigma_1^2) - (y-\mu)^2/(2\sigma_2^2) = 0,$$

where x and y are instances of X and Y. This equation can be rewritten as:

$$(x-\mu)^2/\sigma_1^2 = (y-\mu)^2/\sigma_2^2;$$

solving this equation with respect to μ implies

$$\text{MLE}(\mu) = (w_1x+w_2y)/(w_1+w_2),$$

where $w_1 = 1/\sigma_1^2$ and $w_2 = 1/\sigma_2^2$. The MLE of μ can further be written as

$$\text{MLE}(\mu) = ux+(1-u)y,$$

where $u = w_1/(w_1+w_2)$. The result is a weighted sum of the observed x and y, where the greater the variance of x, the smaller the weight of x.

CMH Test Statistic

A problem associated with directly weighting the sample data based on a confounding variable is that unnecessary stratification can lead to a decrease in the precision. The CMH (Cochran-Mantel-Haenszel) test is used to determine if there is a significant association between two variables while controlling for a confounding variable. Suppose that after stratification, the raw data (one contingency table) is separated into n tables. Define the sample odds ratio of the i^{th} contingency table θ_i. In the i^{th} table below

a	b
c	d

,

each odds ratio θ_i is the odds of success for category X (a/b) over that of Y (c/d), which is (ad)/(bc).

The null hypothesis for CMH test is $\theta_1 = ... = \theta_n = 1$, with an alternative hypothesis: at least one θ is not equal to 1. In other words, the CMH test tests whether the tables have different odds ratios, or whether the stratification is necessary.

The CMH test estimate for the total population odds ratio is the weighted average of the θ's $\frac{\sum_k u_k\theta_k}{\sum_k u_k}$, where each weight of the k^{th} contingency table $u_k = (b_uc_u)/(a+b+c+d)$, representing the inverse of variances from hypergeometric distributions. The estimate simplifies to:

$$\frac{\sum_k a_kd_k/(a_k+b_k+c_k+d_k)}{\sum_k b_kc_k/(a_k+b_k+c_k+d_k)}.$$

The statistic uses the intuition behind the weighting of stratified data.

The CMH test statistic is derived by conditioning on the marginal distributions just like Fisher's exact test, thus resulting in n hypergeometric distributions, but only leaving the first cell of each contingency table free. This is because the if the frequency of each column and row is being fixed, or being conditioned on, then the odds ratio of the k^{th} contingency table can be expressed as a function of a_k only. Specifically, the statistic is $\frac{[\sum_k (a_k - \mathrm{E}(a_k))]^2}{\sum_k Var(a_k)}$, where:

$$\mathrm{E}(a_k) = (a_k+b_k)(a_k+c_k)/(a_k+b_k+c_k+d_k);$$

$$Var(a_k) = (a_k+b_k)(c_k+d_k)(a_k+c_k)(b_k+d_k)/ [(a_k+b_k+c_k+d_k)^2(a_k+b_k+c_k+d_k-1)].$$

Under the null hypothesis that the odds ratios are all 1, and when the sample size is large enough, the statistic follows a chi-squared distribution with 1 degree of freedom, since the chi-squared test can be used to determine whether there is a difference between the null distribution of frequencies and the observed distribution of frequencies, and 2-by-2 contingency tables are used, so the degrees of freedom is (2-1)(2-1) = 1. By using the CMH test rather than a series of chi-squared tests, we can lower the Type I error rate by avoiding the multiple comparison problem, which arises when we perform multiple statistical tests on the same data set. When we perform multiple tests, the probability of observing at least one significant result by chance increases., thus increasing the occurrence of Type I errors. ■

William Haenszel

Nathan Mantel

BAYESIAN STATISTICS

Thomas Bayes

Thomas Bayes (1702-1761) was an English mathematician, statistician, and Presbyterian minister, who is best known for developing Bayes theorem. Bayes theorem is a fundamental concept in probability theory that describes how to update the probability of a hypothesis based on new evidence. Today, Bayes' theorem is used in a wide range of fields, including statistics, computer science, economics, and engineering, among others.

Interested?
Contact the author at sophia.yx.zhu@gmail.com.

DIAGNOSTIC RATIO LIKELIHOOD

Proof of the Bayes Theorem

The Bayes theorem states:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} = \frac{P(X|Y)P(Y)}{\int_{-\infty}^{\infty} P(Y)P(X|Y)d(Y)}.$$

First, we can derive the following:

$$P(Y|X) = \frac{P(X \cap Y)}{P(X)}$$

directly from the fundamental rule of probability, and P(X Y) can further be written as:

$$P(X \cap Y) = P(X)P(Y|X)$$

using the multiplication rule of probability.

We can then expand the denominator using the total probability theorem. That is,

$$P(X) = \sum_{i=1}^{n} P(X \cup Y_k) = \sum_{i=1}^{n} P(Y_k)P(X|Y_k)$$

if X is discrete or

$$P(X) = \int_{-\infty}^{\infty} P(Y)P(X|Y)d(Y)$$

if X is continuous. The probability $P(X)$ is also known as the marginal likelihood of X.

Diagnostic Ratio Likelihood

The diagnostic likelihood ratio (DLR) is a statistical measure that quantifies how much the results of a diagnostic test can change the probability of a patient having a particular condition. To show how the post-test odds are calculated, We first define the following metrics:

1) The sensitivity is the probability of the test being positive given the subject has the disease, or $P(+ \mid D)$.

2) The specificity is the probability of the test being negative given the subject doesn't have the disease, or $P(- \mid D^C)$.

3) DLR of a positive test (DLR^+) is the odds of correctly outputting a positive result, which is $P(+ \mid D) / P(+ \mid D^C)$, or sensitivity / (1 - specificity).

4) DLR of a negative test (DLR^-) is the odds of correctly outputting a negative result, which is $P(+ \mid D) / P(+ \mid D^C)$, or (1 - sensitivity) / specificity.

5) The positive pre-test odds refer to the odds of disease before the test, or $P(D) / P(D^C)$. Similarly, the negative pre-test odds are $P(D^C) / P(D)$.

6) The positive post-test odds refer to the odds of disease after the test, or $P(D \mid +) / P(D^C \mid +)$. Similarly, the negative post-test odds are $P(D \mid -) / P(D^C \mid -)$.

From the discrete version of the Bayes rule, we have:

$$P(D|+) = \frac{P(+|D)P(D)}{P(+|D)P(D)+P(+|D^C)P(D^C)}$$

and

$$P(D^C|+) = \frac{P(+|D^C)P(D^C)}{P(+|D)P(D)+P(+|D^C)P(D^C)}.$$

Dividing $P(D \mid +)$ by $P(D^C \mid +)$, we get

$$\frac{P(D|+)}{P(D^C|+)} = \frac{P(+|D)}{P(+|D^C)}\frac{P(D)}{P(D^C)}.$$

So, the positive post-test odds are the DLR^+ times the positive pre-test odds. The result similarly applies to the negative post-test odds. ■

CREDIBLE INTERVAL AND THE BETA DISTRIBUTION

Credible Intervals *vs.* Confidence Intervals

Bayesian credible intervals are a way of quantifying the uncertainty in an estimated parameter or set of parameters in a Bayesian statistical model. Unlike frequentist confidence intervals, which are based on repeated sampling and probability limits, Bayesian credible intervals are based on the posterior distribution of the parameter(s) of interest, which takes into account both the data and any prior information.

Constructing a Credible Interval

To construct a Bayesian credible interval, we first need to define a prior distribution that represents our beliefs about the parameter of interest before seeing the data. Then, we need to compute the posterior distribution of the parameter given the observed data. Finally, we can use the posterior distribution to compute the credible interval, which is a range of values that contains a specified percentage of the posterior distribution.

For example, if we want to con-struct a 95% credible interval for the mean of a normal distribution, we could start with a prior distribution for the mean (e.g., a normal distribution with a specified mean and variance). Then, we could use Bayes' theorem to update the prior to the posterior distribution, which would be another normal distribution with updated mean and variance based on the observed data. Finally, we could compute the 2.5^{th} and 97.5^{th} percentiles of the posterior distribution to obtain the 95% credible interval. This interval would contain the true mean with 95% probability, given the observed data and the prior distribution.

Strength

Credible intervals are a strength of Bayesian statistics because they provide a probabilistic statement about the range of plausible values for a parameter, given the observed data and the prior distribution. The interval can be interpreted as the range of values for the parameter that has a high probability of being the true value, given the data and the prior.

The credible interval can also be used to make decisions with regard to the null hypothesis rather than the alternative hypothesis. If the null hypothesis falls within the credible interval, then it is

Jakob Bernoulli

Jakob Bernoulli contributed to the development of calculus, probability theory, the mathematical constant e, and the binomial distribution.

plausible that the null hypothesis is true, given the data and the prior. If the null hypothesis does not fall within the credible interval, then it is less plausible that the null hypothesis is true, given the data and the prior.

However, it is important to note that the strength of the credible interval depends on the prior distribution chosen. If the prior distribution is not well-informed or is misspecified, then the resulting credible interval may not accurately reflect the range of plausible values for the parameter. Therefore, careful consideration should be given to the choice of prior distribution in Bayesian analysis.

The Beta Distribution

The beta distribution is a continuous probability distribution that is defined on the interval [0, 1]. It is a two-parameter family of distributions, and its shape is determined by the values of these parameters. The distribution is often used as a model for the distribution of probabilities or proportions.

The probability density function (PDF) of the beta distribution is given by:

$$f(x) = \frac{1}{B(a,b)} \cdot x^{a-1} \cdot (1-x)^{b-1}.$$

where x is a random variable between 0 and 1, a and b are positive shape parameters, and $B(a, b)$ is the beta function, defined as:

$$B(a,b) = \int_0^1 t^{a-1} \cdot (1-t)^{b-1} dt.$$

The reciprocal of the beta function serves as an adjustment to the density function, scaling its cumulative probability to 1.

The expected value of the beta distribution is $\mathrm{E}[x] = \frac{a}{a+b}$, and the variance is $\mathrm{Var}[x] = \frac{ab}{(a+b)^2 \cdot (a+b+1)}$. The beta distribution is often used as a conjugate prior for the binomial distribution in Bayesian analysis, as it allows for closed-form updates of the posterior distribution. The beta distribution is also used in a wide range of applications, including modeling proportions, Bayesian inference, and Bayesian hypothesis testing.

By using a beta distribution as the conjugate prior of a binomial Bayesian model, we observe that the posterior is also a beta distribution. Specifically, if we have a binomial distribution with parameters n and p, and we use a beta distribution with parameters a and b as the prior for p, then the posterior distribution for p given the data is also a beta distribution with updated parameters a' and b', where:

a' = a + number of successes in the data;
b' = b + number of failures in the data.

To see the above, we first write out $P(x = k|p) = \binom{n}{k} p^k (1-p)^{n-k}$. Using Bayes theorem, $P(p|k) \propto P(k|p)P(p)$. Substituting the beta distribution as the prior, we have $P(p|k) \propto (p^k(1-p)^{n-k})(p^{a-1}(1-p)^{b-1})$. This implies $P(p|k) \propto p^{k+a-1}(1-p)^{n-k+b-1}$. This is equivalent to $p^{a'}(1\text{-}p)^{b'}$. In the resulting posterior distribution, there should be a coefficient of $\binom{n}{k}$ and a scaling factor defined by the reciprocal of the beta function. Thus, we have shown that the beta prior yields a Bayesian model that also follows a beta distribution, which is called the beta-binomial distribution.

Nuisance Parameters

In statistics, a nuisance parameter is a parameter that is not of direct interest in a study but must be accounted for in order to make valid inferences about the parameters of interest.

In Bayesian statistics, nuisance parameters are typically treated as random variables and marginalized out of the posterior distribution. This means that credible intervals take into account the uncertainty in the nuisance parameters and reflect the uncertainty in the parameter of interest, given the data and any prior information, making Bayesian statistics particularly useful in modeling the variation/noises in machine learning.

The process of eliminating nuisance parameters is called marginalization, and it involves integrating the joint posterior distribution over the nuisance parameters to obtain the marginal posterior distribution of the parameters of interest. This can be done using Bayes' theorem and the law of total probability, which states that the joint probability of two or more events is equal to the sum of their individual probabilities.

For example, suppose we have a Bayesian model with two parameters of interest, θ and φ, and a nuisance parameter, Γ. Let y denote the posterior. The joint posterior distribution of the three parameters can be written using the Bayes rule:

$$P(\theta, \phi, \Gamma|y) \propto P(y|\theta, \phi, \Gamma)\pi(\theta)\pi(\phi)\pi(\Gamma).$$

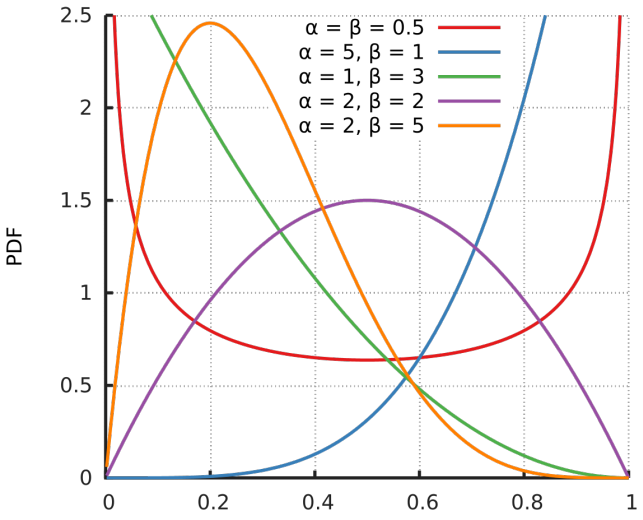
To eliminate the nuisance parameter gamma, we can integrate over all possible values of Γ, which gives us the marginal posterior distribution of θ and φ:

$$p(\Gamma|y) = \int_{-\infty}^{\infty} P(\theta, \phi, \Gamma|y)d(\Gamma) \propto \int_{-\infty}^{\infty} P(y|\theta, \phi, \Gamma)\pi(\theta)\pi(\phi)\pi(\Gamma)d(\Gamma).$$

This marginal posterior distribution of theta and phi represents our updated beliefs about these parameters based on the data, after marginalizing out the nuisance parameter Γ.

In contrast, frequentist statistics typically treat nuisance parameters as fixed, unknown values and use various techniques such as maximum likelihood estimation or method of moments to estimate their values. Confidence intervals are then constructed to reflect the uncertainty in the parameter of interest, given the data and the estimated values of the nuisance parameters.

In practice, marginalization can be computationally expensive. However, eliminating nuisance parameters can lead to more accurate inference and more informative posterior distributions for the parameters of interest. In general, Bayesian methods can be more flexible in dealing with nuisance parameters. ■



LINEAR ALGEBRA

THE FIBONACCI SEQU--ENCE & LINEAR MAPS

Formulating the Problem

Let $f_1, f_2, \ldots$ denote the Fibonacci sequence for each f_i in **F** defined by

$f_1 = 1, f_2 = 1$, and $f_n = f_{n-1} + f_{n-2}$.

The problem is to find the value of the terms in the Fibonacci sequence.

Defining a Linear Map

We first define an operator T on $\mathbf{R}^2$ as $T(x, y) = (y, x+y)$. We easily get that $T^n(0, 1) = T \bullet T \bullet \ldots \bullet T(0, 1) = T \bullet T \bullet \ldots \bullet T(1, 1) = \ldots = (f_n, f_{n+1})$ by definition of T.

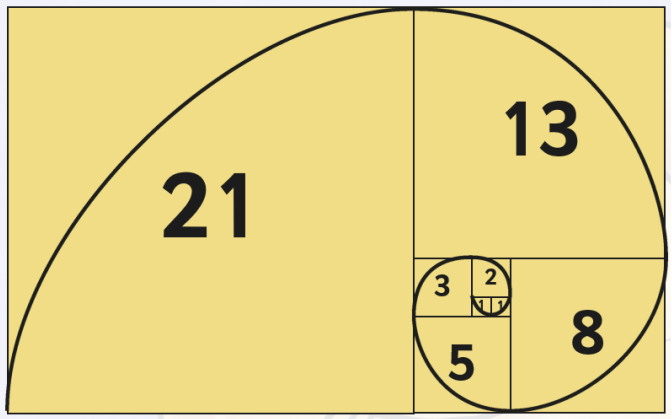
The eigenvalues of T are found through the set of equations

$$\begin{cases} x = \lambda y \\ y = \lambda(x + y) \end{cases}$$

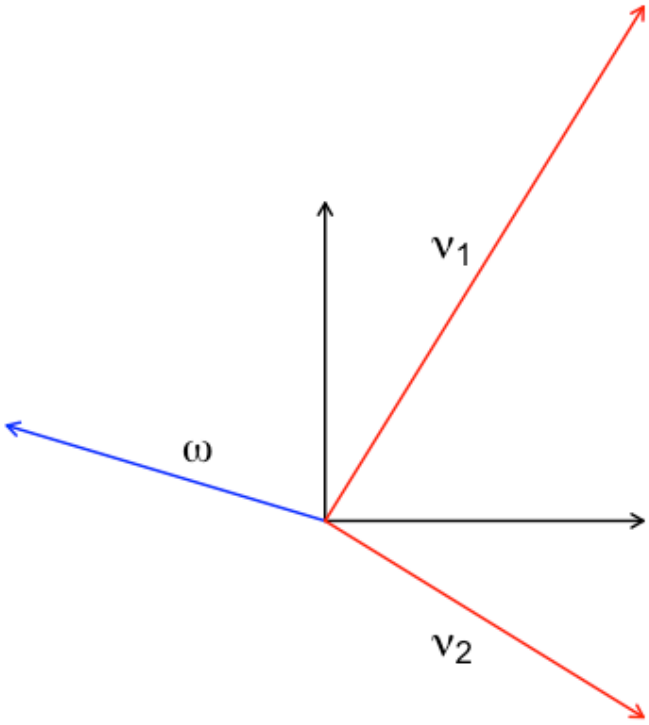
as $\frac{1+\sqrt{5}}{2}$ and $\frac{1-\sqrt{5}}{2}$. The eigenvectors are naturally of the form $(x, \frac{1+\sqrt{5}}{2}x)$ and $(x, \frac{1-\sqrt{5}}{2}x)$ by $x = \lambda y$.

We can write a new set of basis for $\mathbf{R}^2$: $b_1 = (1, \frac{1+\sqrt{5}}{2})$ and $b_2 = (1, \frac{1-\sqrt{5}}{2})$. After the change in basis, $(0, 1) = \frac{1}{\sqrt{5}}(b_1 - b_2)$. Thus, $T^n(0, 1) = \frac{1}{\sqrt{5}}T^n(b_1 - b_2) = \frac{1}{\sqrt{5}}(T^n b_1 - T^n b_2) = (f_n, f_{n+1})$.

$T^n(b_1) = T \bullet T \bullet \ldots \bullet T(1, \frac{1+\sqrt{5}}{2}) = T \bullet T \bullet \ldots \bullet T(\frac{1+\sqrt{5}}{2}, (\frac{1+\sqrt{5}}{2})^2) = (\frac{1+\sqrt{5}}{2})^n b_1$. Similarly, $T^n(b_2) = (\frac{1-\sqrt{5}}{2})^n b_2$. Hence, $\frac{1}{\sqrt{5}}(T^n b_1 - T^n b_2) = \frac{1}{\sqrt{5}}((\frac{1+\sqrt{5}}{2})^n b_1 - (\frac{1-\sqrt{5}}{2})^n b_2) = (f_n, f_{n+1})$, which implies $f_n = \frac{1}{\sqrt{5}}((\frac{1+\sqrt{5}}{2})^n - (\frac{1-\sqrt{5}}{2})^n)$, the expression for the n^{th} term in Fibonacci. ■

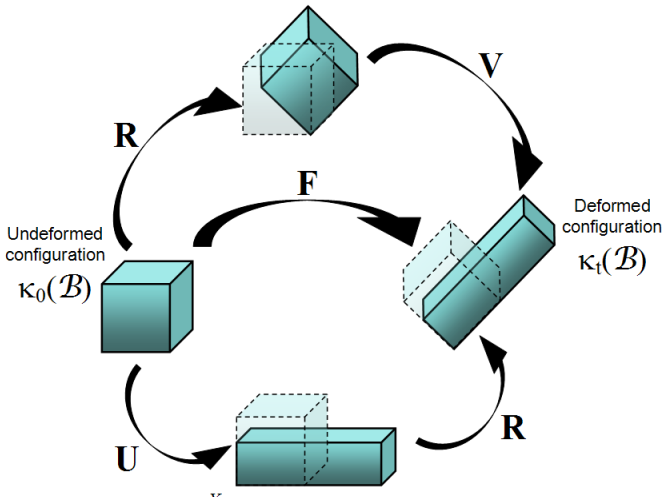


Geometric Representation of the Fibonacci Sequence



Two Sets of Basis:
We switched from the standard (0, 1) and (1, 0) set to b_1 and b_2, changing the relative axis of vectors in $\mathbf{R}^2$.

POLAR DECOMPOSITI-ON OF LINEAR MAPS



Complex Numbers

Recall that any v in $\mathbf{C}$ can be expressed as $|v|e^{i\theta}$, where $|v| = \sqrt{\bar{v}v}$ is a scaling coefficient representing the length of v and $e^{i\theta}$ is a complex number of unit length representing the direction of v on the complex coordinate. It can be seen as decomposing v into a scaling $|v|$ and a rotation by θ.

Polar Decomposition

The polar decomposition theorem states that any linear map A in $\mathbf{C}^{n\times m}$ can be decomposed into a product of two matrices, one of which is a unitary isometric matrix and the other is a positive-semidefinite Hermitian matrix. The theorem states that $A = UP$, where U in $\mathbf{C}^{n\times m}$ is a unitary matrix, which means that its inverse is equal to its conjugate transpose, and P in $\mathbf{C}^{m\times m}$ is a positive-semidefinite Hermitian matrix, which means that all of its eigenvalues are non-negative.

Similar to the polar form of complex numbers, we can first derive the value of P, representing the scaling factor of A:

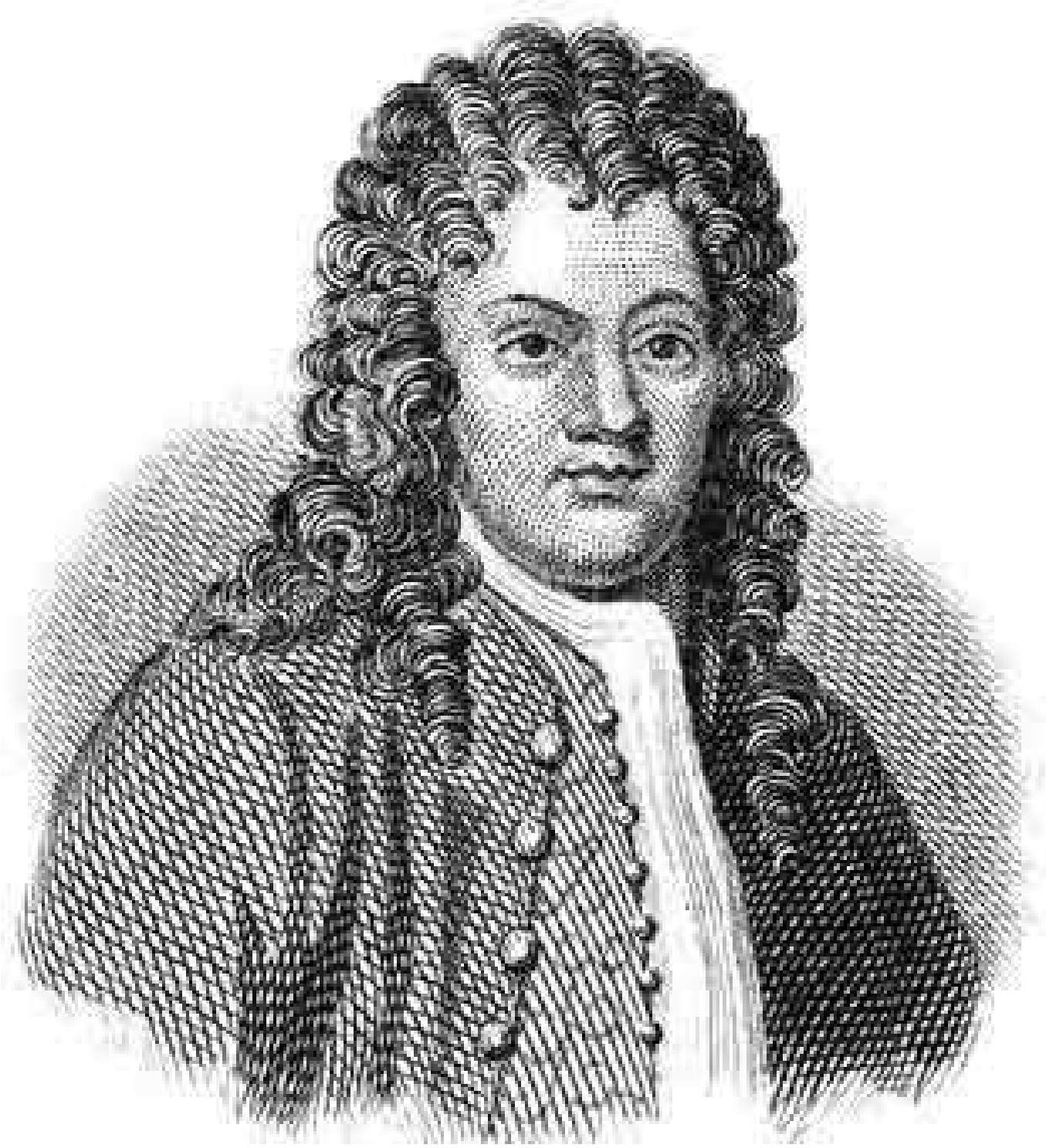
$$A^TA = (UP)^{\mathrm{T}}(UP) = P^T(U^TU)P.$$

Since U is the rotational factor of A, or a rotation matrix, it can be shown that the columns of U with regard to the standard bases of an m-dimensional vector space are orthogonal to each other. Then, $U^TU = \mathrm{I}_m = U^{-1}U$. This implies $A^TA = P^T(U^TU)P = P^TP$, so $P = (A^TA)^{1/2}$.

On the other hand, the rotational factor U can be found by substituting P into $A = UP$, implying $U = A(A^TA)^{-1/2}$. In other words, let $\lambda_1, ..., \lambda_n$ be the eigenvalues of $(A^TA)^{1/2}$ and $x_1, ..., x_n$ be the eigenvectors. There exists an integer $r \leq n$ such that $\lambda_{r+1}, ..., \lambda_n = 0$ as A is not necessarily full rank and invertible. We establish that $Px_i = \lambda_i x_i$. Since $Ax_i = UPx_i = U\lambda_i x_i$, implying $Ux_i = (1/\lambda_i)Ax_i$ when $i \leq r$ and $Ux_i = x_i$ when $i > r$. U can be written as:

$$\begin{bmatrix} \frac{1}{\lambda_1}Ax_1 & \dots & \frac{1}{\lambda_r}Ax_r & x_{r+1} & \dots & x_n \end{bmatrix} \begin{bmatrix} x_1^H \\ \vdots \\ x_n^H \end{bmatrix},$$

which also suffices the previous definition that $U = A(A^TA)^{-1/2}$. ■



Brook Taylor (1685-1731) was an English mathematician who made significant contributions to several areas of mathematics, including calculus, algebra, and geometry. He is best known for his work on Taylor series, which provide a way to represent functions as infinite sums of polynomial terms.

DERIVATION OF THE TAYLOR SERIES USING VECTOR PROJECTION

Projection

The Taylor series can be seen as the orthogonal projection of a continuous function f in **V** onto a polynomial g in polynomial space **P**. Specifically, we want to minimize the quantity $\|f - g\|^2 = \langle f, g \rangle = \int f(x)g(x)\,dx$.

The shortest distance between the f and g is given by the vertical projection of f onto g, since obviously $\|f - \text{proj}_g f\|^2 \leq \|f - \text{proj}_g f\|^2 + \|\text{proj}_g f - g\|^2 = \|f - g\|^2$ by the Pythagorean theorem. To minimize the *L2* loss $\|f - g\|^2$, the optimal value of g is given by $(\text{proj}_{\mathbf{P}} f)$. The projection of f onto the polynomial space **P** can be decomposed into:

$$g = \text{proj}_{\mathbf{P}} f = \langle b_1, f \rangle b_1 + \ldots + \langle b_n, f \rangle bn$$

with regard to an orthonormal basis set $(b_1, \ldots, b_n)$.

Finding the Basis

The remaining task is to identify the orthonormal basis. Using the Gram-Schmidt orthonormalization process, given a basis set $(e_1, \ldots, e_n)$, we can find the b's as follows:

$$b_1 = e_1.$$

$$b_2' = (e_2 - \text{proj}_{\text{span}(b_1)} e_2);\ b_2 = b_2' / \|b_2'\|.$$

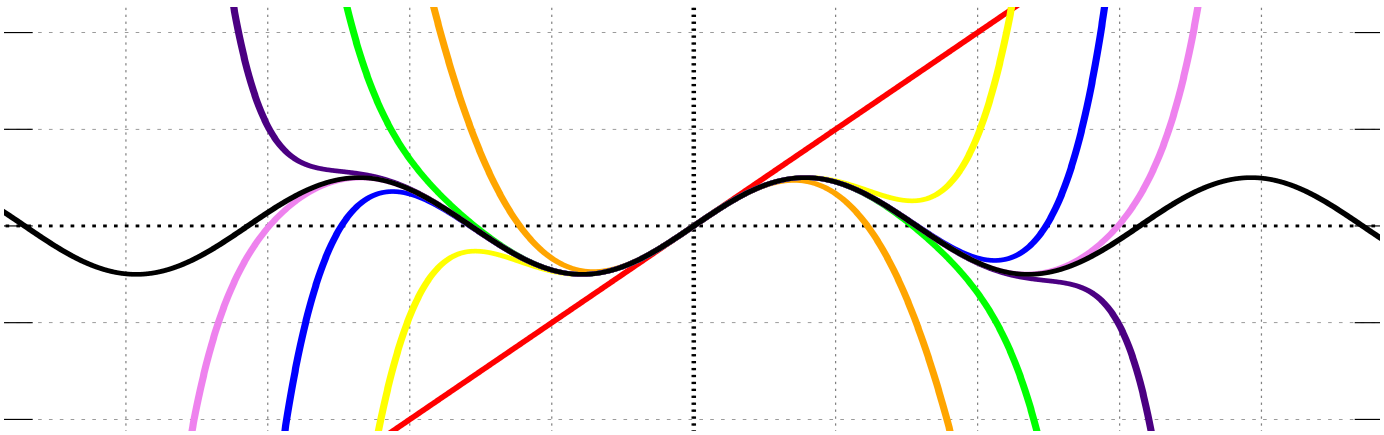
$$\ldots$$

$$b_n' = e_n - \text{proj}_{\text{span}(b_1)} e_n - \ldots - \text{proj}_{\text{span}(b_{n-1})} e_n;$$

$$b_n = b_n' / \|b_n'\|.$$

This process eliminates the components of the basis vector that have already been accounted for by previous basis vectors, and then normalizes them, where each b_i' denote an intermediate step that has not yet been normalized yet.

The result would be the same as the Taylor series. ■



ALTERNATE DEFINITIO-N OF AFFINE SUBSETS

Definition

An affine subset W of a vector space V is x+U, for x in V and U a subspace of V. An alternate definition of affine subsets is W is an affine subset if and only if for every v and w in W, $(1-\lambda)v + \lambda w$ is in W.

Proof

If W is an affine subset, W = x+U for chosen v and U. Let v and w be arbitrary vectors in W. Then, $v = x+u_i$ and $w = x+u_j$ for chosen u_i, u_j in U. It follows that:

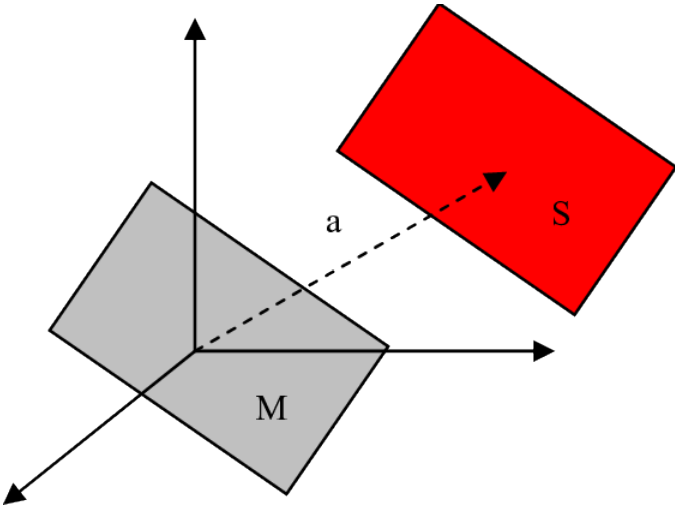
$$\begin{aligned}&(1-\lambda)v + \lambda w\\ &= (1-\lambda)(x+u_i) + \lambda(x+u_j)\\ &= x + u_i - \lambda x - \lambda u_i + \lambda x + \lambda u_j\\ &= x + u_i - \lambda u_i + \lambda u_j.\end{aligned}$$

By definition of a subspace, $(u_i - \lambda u_i + \lambda u_j)$ must be in U. Thus, $(1-\lambda)v + \lambda w = x + u_i - \lambda u_i + \lambda u_j$ must be in W.

Conversely, if $(1-\lambda)v + \lambda w$ is in W, it is implied that $v + \lambda(w - v)$ is in W. This further implies $\lambda(w - v)$ is in the subset (W - v). On the other hand, since w is in W, $(w - v)$ is in the subset (W - v). Hence, if v and w are in W, $(1-\lambda)v + \lambda w$ implies that W is closed under scalar multiplication.

Let v' and w' also be vectors in W. Then, $(w' - v)$ and $(v' - v)$ is in the subset (W - v). Since $(1-\lambda)v' + \lambda w'$ is in A, taking λ=0.5 gives us $0.5(v' + w')$ is in W, or $(0.5v' + 0.5w' - v)$ is in (W - v). By closure of scalar multiplication, $(v' + w' - 2v) = (v' - v) + (w' - v)$ is in the subset (W - v). Therefore, for arbitrary vectors $(w' - v)$ and $(v' - v)$ in (W - v), $(v' - v) + (w' - v)$ is also in (W - v). In other words, (W - v) is closed under vector addition. Since v is in W, it can be further inferred that W is closed under addition.

Since W is closed under scalar multiplication and vector addition, it is an affine subset of V. ■



An affine subset W of a vector space V is a subset closed under scalar multiplication and vector addition. It does not contain the origin but is parallel to a subspace U of the V, as shown in the picture as the red plane parallel to the grey one.

MACHINE LEARNING

VARIANCE & EXPLODING OR VANISHING GRADIENT

Parameter Initialization

In deep learning, when computing the gradients in the backward pass, sometimes vanishing gradients or exploding gradients are encountered. Vanishing gradients are when the gradients become extremely small as the weights that are multiplied to the gradient signals get very small. In contrast, exploding gradients are when the gradients become extraordinarily large. If the parameters are initialized poorly, the model may have difficulty converging to an optimal solution during training. Hence, parameter initialization becomes especially important. In this article, we will explore how parameters in the weight matrices Ω_k at the k^{th} layer can be initialized to best prevent inefficient training (e.g., vanishing or exploding gradients) and poor performing models with a ReLU activation function.

Variance of Standard ReLU

The ReLU activation function follows:

$$X' = \text{ReLU}(X) = \begin{cases} X, & \text{if} X \geq 0; \\ 0, & \text{if} X < 0. \end{cases}$$

Suppose we initialize the bias vectors β_k as zero vectors and the Ω_k's using Gaussian with mean (μ) 0. Then we would want the weights to be evenly spread out so that none of the weights are extremely large or small. Therefore, the task is to find the optimal variance of Ω_k.

Let h_k be the k^{th} hidden layer, f_k be $\beta_k + \Omega_k h_k$, and D_k be the dimension of h_k. h_k is naturally $\text{ReLU}(f_{k-1}) = \text{ReLU}(\beta_k + \Omega_k h_{k-1})$. It follows that the variance of each i^{th} element of f_k,

$$\begin{aligned} &\text{Var}(f_{i,k}) \\ &= \text{E}(f_{i,k}^2) - \text{E}(f_{i,k})^2 \\ &= \text{E}([\beta_{i,k} + \Sigma_j \, \Omega_{(i,j),k} h_{j,k}]^2), \end{aligned}$$

as $\text{E}(f_k) = \text{E}(\beta_k) + \text{E}(\Omega_k h_k) = 0 + \text{E}(\Omega_k)\text{E}(h_k) = 0$, and the summation of j from 0 to D_k is derived from matrix multiplication. Thus,

$$\begin{aligned} &\text{Var}(f_k) \\ &= \text{E}(\beta_{i,k}^2 + \Sigma_j \, 2\beta_{i,k}\Omega_{(i,j),k} h_{j,k} + \Sigma_j \, \Omega_{(i,j),k}^2 h_{j,k}^2) \\ &= \text{E}(\Sigma_j \, \Omega_{(i,j),k}^2 h_{j,k}^2) \\ &= \Sigma_j \, \text{E}(\Omega_{(i,j),k}^2)\text{E}(h_{j,k}^2), \end{aligned}$$

as β_k is a zero vector, so $\beta_{i,k} = 0$. Since

$$\begin{aligned} &\Sigma_j \, \text{E}(\Omega_{(i,j),k}^2) \\ &= Var(\Omega_{(i,j),k}) - \text{E}(\Omega_{(i,j),k})^2 \\ &= Var(\Omega_{(i,j),k}) \text{ and } \text{E}(h_{j,k}^2) \\ &= Var(h_{j,k}) - \text{E}(h_{j,k})^2 \\ &= Var(h_{j,k}), \Sigma_j \, \text{E}(\Omega_{(i,j),k}^2)\text{E}(h_{j,k}^2) \\ &= \Sigma_j \, Var(\Omega_{(i,j),k}) Var(h_{j,k}) \\ &= D_k Var(\Omega_{i,k}) Var(h_k) \\ &= D_k Var(\Omega_{i,k})\text{E}(h_k^2). \end{aligned}$$

Summing up every i^{th} element in f_k, the total variance of f_k is $D_k Var(\Omega_k)\text{E}(h_k^2)$.

Now, we will show that

$E(h_k^2) = 0.5Var(f_{k-1})$.

First, $Var(f_{k-1}) = E(f_{k-1}^2) - E(f_{k-1})^2 = E(f_{k-1}^2) - 0 = \int_{-\infty}^{\infty} f_{k-1}^2 \cdot p(f_{k-1})d(f_{k-1})$

$$E(h_k^2) = \int_{-\infty}^{\infty} \max(0, f_{k-1}^2) \cdot p(f_{k-1})d(f_{k-1})$$
$$= \int_{0}^{\infty} f_{k-1}^2 \cdot p(f_{k-1})d(f_{k-1})$$
$$= \frac{1}{2}\int_{-\infty}^{\infty} f_{k-1}^2 \cdot p(f_{k-1})d(f_{k-1}) = \frac{1}{2}Var(f_{k-1})$$

Substituting the above result into the $Var(f_k)$, we get $Var(f_k) = 0.5D_kVar(\Omega_k)Var(f_{k-1})$. Since our goal is to maintain the variance across layers (making $Var(f_{k-1})$ and $Var(f_k)$ as close as possible), we should set $Var(\Omega_k)$ to be $2/D_k$.

Similarly, in the backward pass, we calculate the variances in reverse order. Thus, we would want $Var(\Omega_k) = 2/D_{k+1}$.

The **He initialization** calculates the $Var(\Omega_k)$ by taking the average of the two variances. That is,

$$Var(\Omega_k) = 0.5(D_k + D_{k+1}).$$

The He initialization helps mitigate the issues of vanishing or exploding gradients by ensuring that the weights are initialized to reasonable values that do not cause the gradients to explode or vanish during backpropagation. This, in turn, can lead to faster and more stable training of deep neural networks and lead to better performance.

Kaiming He, Chinese computer scientist and AI researcher, and the inventor of several influential algorithms, including ResNet and He initialization

BAYESIAN LOSS IN IMAGE RESTORATION

The Bayesian Framework

Image denoising is a process of removing noise from an image. Bayesian image denoising is a popular method that involves the use of a Bayesian framework to estimate a clean image from a noisy observation.

In Bayesian image denoising, the goal is to find the posterior probability distribution of the clean image given the noisy observation. This is given by Bayes' theorem: $P(Im_{clean}|Im_{noisy}) = \frac{P(Im_{noisy}|Im_{clean}) \cdot P(Im_{clean})}{P(Im_{noisy})}$, where $P(Im_{clean}|Im_{noisy})$ is the posterior distribution of the clean image, $P(Im_{noisy}|Im_{clean})$ is the likelihood of the noisy observation given the clean image, $P(Im_{clean})$ is the prior distribution of the clean image, and $P(Im_{noisy})$ is the evidence term that normalizes the posterior distribution.

To estimate the clean image, a Bayesian loss function is used. The Bayesian loss function is defined as the negative logarithm of the posterior distribution, which measures the discrepancy between the estimated and the true clean image.

By minimizing the Bayesian loss function, we can find the optimal estimate of the clean image that is most consistent with the noisy observation and the prior distribution.

One of the advantages of the Bayesian approach is that it allows us to incorporate prior knowledge about the image into the denoising process, such as the smoothness or sparsity of the image. This helps to reduce the impact of noise on the image while preserving its important features.

Noise To Void (N2V)

The N2V algorithm (Krull et al.) was developed to resolve the impractical limitations of the Noise To Noise (N2N) algorithm (Lehtinen et al.), which requires two versions of noisy images of the same clean image so that they can learn from each other the underlying distribution of the clean image. N2V, on the other hand, requires only one noisy image to recover the clean image.

The basic idea behind N2V is to train a deep neural network to predict the original image pixels from the noisy input image pixels. The network is trained using a Bayesian loss function that incorporates the properties of the noise distribution. The N2V algorithm uses a U-Net network architecture with masked blind spots, where the encoder part of the

network maps the noisy image to a low-dimensional feature space, and the decoder part of the network maps the low-dimensional features back to the denoised image. Before we pass the image through the DNN, we first mask the noisy image through masks with certain shapes that generate blindspots, representing potential noisy pixels that need restoration using the distribution of adjacent pixels.

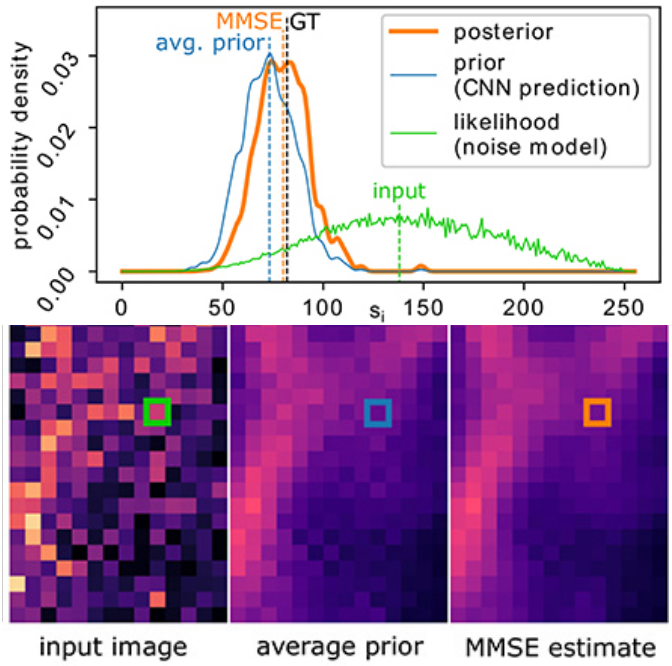
The N2V algorithm uses a Bayesian loss function, which involves defining a prior distribution over the clean image space and then finding the maximum of the posterior distribution given the noisy input image. Denote the environment around a pixel, or its surrounding pixels, as Env_{pixel}. We obtain the following expression directly from the Bayes rule: $P(Im_{clean}|Im_{noisy}) \propto P(Im_{noisy}|Im_{clean})P(Im_{clean}|Env_{pixel})$. In the above equation, the posterior distribution of the clean image given the noisy image is unknown, while the prior distribution of the noisy image given the clean image is assumed to be a Gaussian distribution with a mean of zero and a variance that depends on the local image structure, and the distribution of the clean image given the environment is the output of the U-Net that we intend to optimize.

Integrating over all potential clean images given the noisy, we get the marginal distribution of the noisy image given the environment using the Bayes rule: $P(Im_{noisy}|Env_{pixel}) = \int_{-\infty}^{\infty} P(Im_{noisy}|Im_{clean})P(Im_{clean}|Env_{pixel})d(Im_{clean})$, where the distribution of the clean images given the environment is the output of the U-Net and the distribution of the noisy image given the environment is observed from the training data.

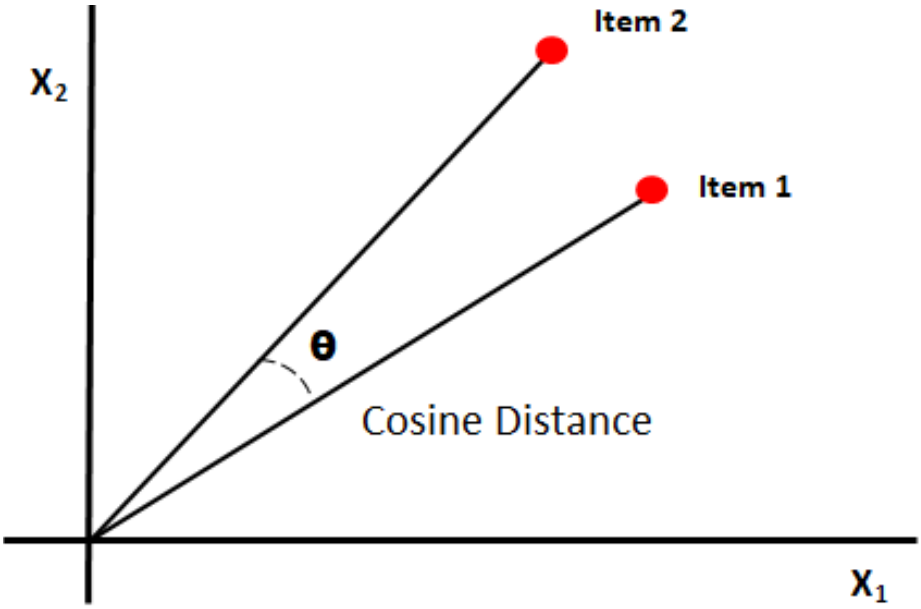
Hence, the likelihood function is also modeled as a Gaussian with a mean equal to the noisy input image and a variance that is a function of the local environment image structure. The formal loss function during training is found by minimizing the negative log-likelihood of the integral, which represents the maximum likelihood estimate of $P(Im_{noisy}|Env_{pixel})$ since such MLE implies the most probable value of the pixel from its context.

During the testing phase, we bring in $P(Im_{clean}|Im_{noisy})$ again from below: $P(Im_{clean}|Im_{noisy}) \propto P(Im_{noisy}|Im_{clean})P(Im_{clean}|Env_{pixel})$. As $P(Im_{clean}|Env_{pixel})$ is obtained from the neural network, we can compute $\mathrm{E}(P(Im_{clean}|Im_{noisy}, Env_{pixel}))$, the predicted clean output.

N2V does not require any explicit ground truth images and can handle different types of noise, including Gaussian, Poisson, and speckle noise. ■



COSINE SIMILARITY LOSS IN NLP



Text Embedding

In natural language processing (NLP), a vector is a mathematical representation of a text document that has each of its word/phrase encoded as a numerical value in a high-dimensional space. This process is called text embedding.

Geometric Interpretation of the Loss

Cosine similarity loss is a measure of the similarity between two vectors, commonly used in natural language processing (NLP). By minimizing the loss, we aim to let two embedded texts represented as vectors be as similar as possible.

The cosine similarity ranges from -1 to 1, with a value of 1 indicating that the vectors are identical, and a value of -1 indicating that they are completely dissimilar. The formula of cosine similarity for two vectors a and b follows:

$$\frac{<\vec{a}, \vec{b}>}{||\vec{a}|| ||\vec{b}||}.$$

Interested in learning more about this magazine?

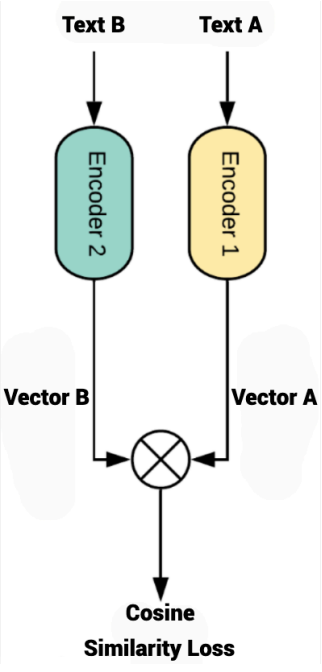
Contact the author at sophia.yx.zhu@gmail.com.

Try This Question:

Proof that if θ is the angle between a and b, $\cos(\theta)$ is equal to:

$$\frac{<\vec{a}, \vec{b}>}{||\vec{a}|| ||\vec{b}||}$$

using the cosine rule. ■



REDUCING ERRORS IN LEAST SQUARES: VARIANCES, BIASES, & NOISES

Dividing Least Squares

The least squares equation between a neural network f with parameter x and the sample outputs $y(x)$ is $L = (f(x) - y(x))^2$. By adding and subtracting a term $\mu(y)$, we get

$$\begin{aligned} & L \\ &= ([f(x) - \mu(y)] + [\mu(y) - y(x)])^2 \\ &= (f(x) - \mu(y))^2 + 2(f(x) - \mu(y))(\mu(y) - y(x)) \\ &+ (\mu(y) - y(x))^2. \end{aligned}$$

Taking the expected value of both sides with respect to y gives

$$\begin{aligned} & E_y(L) \\ &= E_y(f(x) - \mu(y))^2 + 2 \bullet E_y(f(x) - \\ &\mu(y)) \bullet E_y(\mu(y) - y(x)) + E_y(\mu(y) - y(x))^2 \\ &= [f(x) - \mu(y)]^2 + 2 \bullet [f(x) - \mu(y)] \bullet 0 + \sigma_y^2 \\ &= [f(x) - \mu(y)]^2 + \sigma_y^2. \end{aligned}$$

Thus, the expected loss can be separated into two parts: the summed squared deviations of the model output and the mean of y and the variance of the sample training outputs. The additional variance in the sample labels is unavoidable. The term σ_y^2 is called the noise of the model, a type of insurmountable error that resulted from the variance of the training data itself.

Since our neural network depends on the sample training input data i, we denote f as f_i. The term $[f_i(x) - \mu(y)]^2$ in $E(L)$ can further be divided into two other types of errors: variance and bias.

$$\begin{aligned} & [f_i(x) - \mu(y)]^2 \\ &= [(f_i(x) - \mu(f_i)) + (\mu(f_i) - \mu(y))]^2 \\ &= [f_i(x) - \mu(f_i)]^2 + 2 \bullet [f_i(x) - \mu(f_i)] \bullet [\mu(f_i) - \\ &\mu(y)] + [\mu(f_i) - \mu(y)]^2. \end{aligned}$$

Taking the expected value with respect to i gives

$$\begin{aligned} & E_i([f_i(x) - \mu(y)]^2) \\ &= [E_i(f_i(x)) - \mu(f_i)]^2 + 2 \bullet 0 \bullet [\mu(f_i) - \mu(y)] + \\ &[\mu(f_i) - \mu(y)]^2 \\ &= [E_i(f_i(x)) - \mu(f_i)]^2 + [\mu(f_i) - \mu(y)]^2. \end{aligned}$$

Thus, the expected loss with respect to the specific training dataset i can be divided into the variance of the neural network f trained on i, the bias term defined by the summed squared deviation between the mean of the neural network output and the mean of y, and the noise. The bias is also known as the systematic error of the model, as it is the deviation between the average output value and the average predicted value. In other words,

$$\begin{aligned} & E_i(E_y([f_i(x) - \mu(y)]^2)) \\ &= [E_i(f_i(x)) - \mu(f_i)]^2 + [\mu(f_i) - \mu(y)]^2 + \sigma_y^2, \end{aligned}$$

with the first term representing variance, the second bias, and the third noise.

Minimizing Errors

Due to the natural variability of the sample training dataset, the noise component of least squares loss (L2 loss) is unavoidable. However, we can reduce the variance and bias, which are reliant on the neural network model.

To minimize the variance, we can simply gather more sample data to be more certain about the specific output of a given input x; as the number of data points approaches infinity, the predicted output converges to the single true unbiased point estimate.

To minimize the bias, we can increase the capacity of our model by adding more hidden layers or hidden units in each layer, which results in more linear regions that help the model to better resemble a curved surface.

However, note that there might exist a trade-off between the minimization of variance and bias. If we increase the capacity of the neural network, the network might become overfitted on the training dataset and be overly affected by the noises, which are not generalizable to the true distribution of data. Thus, an optimal capacity of the neural network should be carefully selected to avoid such problems. ■